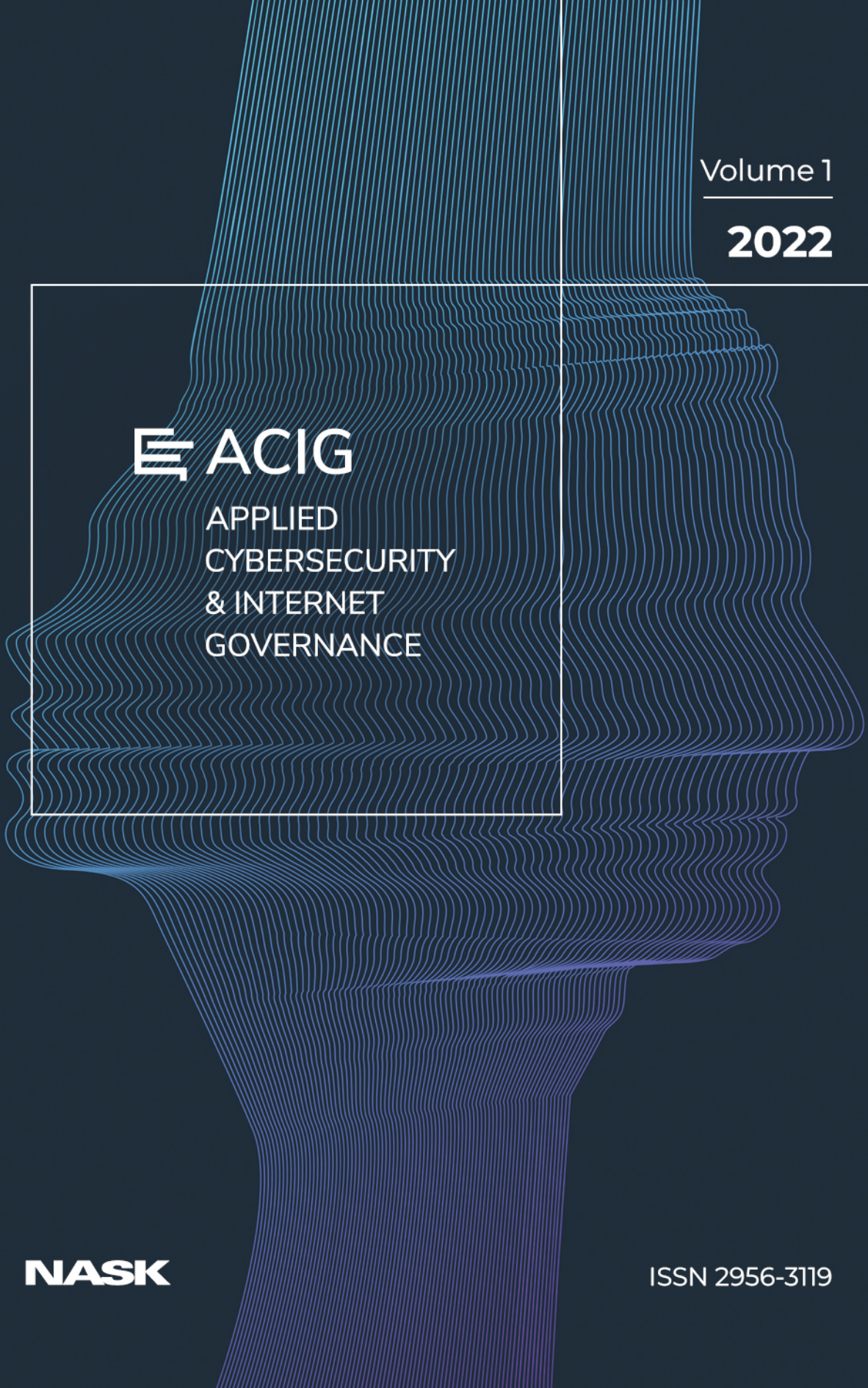


Volume 1

2022

ACIG

APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

ISSN 2956-3119

Volume 1

2022

ACIG

APPLIED
CYBERSECURITY
& INTERNET
GOVERNANCE

NASK

ISSN: 2956-3119 | E-ISSN: 2956-4395

Table of Contents

Letter from the Editor-in-Chief

The world is constantly changing. New, effective security systems require a better understanding of cyberspace and advanced technologies. Traditional cyber defenses cannot keep pace with the high access in 5G networks and IoT delivery, nefarious manipulation of data, malicious disinformation and global corporations. The exposure of cybersecurity vulnerabilities proves the need for more up-to-date scholarly research that will cross borders in technology, policy and application. National Research Institute NASK, a leader in cyber innovations in Poland, is creating a ground-breaking new scientific platform for research publication and multidivisional exchange of ideas in computer science and security studies. To address this gap in knowledge, we are launching a new open-access peer-reviewed scholarly journal Applied Cybersecurity and Internet Governance!

Applied Cybersecurity and Internet Governance responds to contemporary challenges faced by modern civilization. Our ambition is to ensure sustainable technological development, promote crucial technological advances and present new research in machine learning. We are confident that interdisciplinary and multi-sectoral approaches to management of the Internet will enhance harmonious global cooperation.

Be part of this process!

We welcome original research papers that extend the existing knowledge in the field of cyber security, AI and the use of the Internet. Accepted research articles that have gone through a rigorous peer-review process will be published online first. The digital revolution needs your studies!

How to publish with us? Submit your paper with our guidelines on the journal homepage!

Your article can transform knowledge!

Editor-in-Chief
Aleksandra Gasztold

Artificial Intelligence for Cybersecurity: Offensive Tactics, Mitigation Techniques and Future Directions

Erwin Adi | **Deloitte Risk Advisory Pty Ltd, Australia**, ORCID: 0000-0001-7120-1967

Zubair Baig | **Deakin University, Victoria, Australia**, ORCID: 0000-0003-0557-1550

Sherali Zeadally | **College of Communication and Information, University of Kentucky, USA,** ORCID: 0000-0002-5982-8190

Abstract

Cybersecurity has benefitted from Artificial Intelligence (AI) technologies for attack detection. However, recent advances in AI techniques, in tandem with their misuse, have outpaced parallel advancements in cyberattack classification methods that have been achieved through academic and industry-led efforts. We describe the shift in the evolution of AI techniques, and we show how recent AI approaches are effective in helping an adversary attain his/her objectives appertaining to cyberattacks. We also discuss how the current architecture of computer communications enables the development of AI-based adversarial threats against heterogeneous computing platforms and infrastructures.

Keywords

adversarial AI, cyber infrastructures, data analysis, supply chain compromise

Corresponding author:
Erwin Adi; Deloitte Risk Advisory Pty Ltd, Australia;
ORCID: 0000-0001-7120-1967;
eadi@deloitte.com.au

Cite this article as: E. Adi, Z. Baig, S. Zeadally, "Artificial Intelligence for Cybersecurity: Offensive Tactics, Mitigation Techniques and Future Directions", ACIG, vol. 1, no. 1, pp. 6–34, 2022, DOI: 10.5604/01.3001.0016.0800

1. Introduction

Artificial Intelligence has enabled cybersecurity researchers and practitioners alike to design and develop cutting-edge solutions to counter the ever-expanding and increasingly sophisticated types of cyberattack that threaten contemporary computing systems and platforms. Increasing production and marketing of AI-based cybersecurity solutions have set the trend during the past decade [1, 2]. Recent advances in the AI research domain have empowered cybersecurity systems to manage machines autonomously, and to safeguard these by creating rapid defence and reprisals against an adversary, in near real-time [3]. On the other hand, the adversary has also gained significant potency and far outreach in his/her attack strength owing to the same advancements in Artificial Intelligence technology. Though the core attacker steps comprising a data breach, namely vulnerability detection, exploitation, post-exploitation and data theft [1], remain the same, the potential impact of an AI-based system deployed to do so is of increasing concern to all. This is due to the shift from traditional (Fig. 1.) to modern Internet architecture (Fig. 2.).

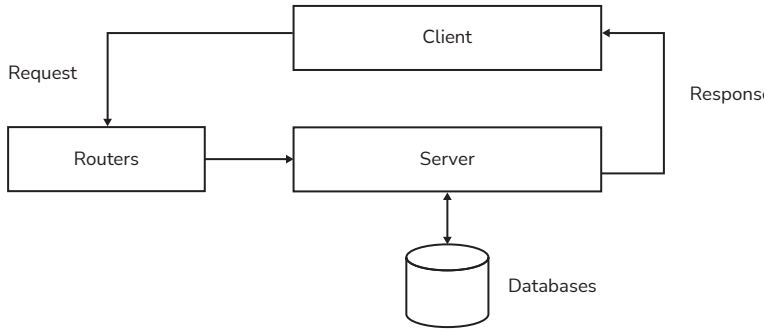


Figure 1. The architecture of the Internet.

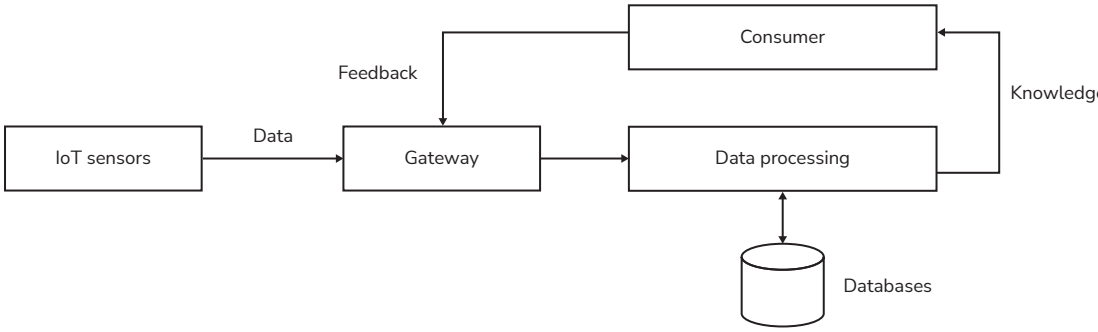


Figure 2. The architecture of the modern Internet, adapted from [4].

The Internet was traditionally viewed as networked interconnections of client-server computers as Fig. 1. shows. The client, such as a PC, sends a request packet to a server. The server processes the request, initiates some actions such as fetching data from the database, and sends its response back to the client. These computers are interconnected by

network devices (e.g. routers). Such a client-server model facilitates the exchange of data, but not knowledge or insightful information that has been processed by an AI machine.

However, the modern Internet needs to be modelled as a communication system not only for exchanging data, but also to render feedback, and knowledge [4] as Fig. 2. shows. A large amount of data is generated by IoT devices, through their sensors that generate phenomena data including user or device location, speech patterns, text, emotions (e.g. through a like button), social links, pictures, and videos. These IoT devices send the data to other devices, including to the Cloud. Their interconnection is served by a gateway that supports heterogeneity in transmission techniques and communication standards. At the other end of the communication line, a machine is responsible for the processing of collected data to improve its usability. It generates outputs from the classification of the data to location recommendations (e.g. a Google map route recommendation). The machine either stores the data in the database or transmits the same (or data converted into knowledge) to a consumer, such as to a smart device, or a monitoring system. This chain of devices collectively comprises an intelligent system, which allows for iterative feeding of data to learn adaptively from sensor data and through device feedbacks.

As a computing device of an intelligent system, each of the resources illustrated in Fig. 2. (i.e. IoT sensors, gateway, data processing, and consumer) can become a target of AI-based cyberattacks (Fig. 3.). They are vulnerable to two attack artifices: those crafted by rational agents or bots, and those that comprise behaviour-mimicry attacks. These two artifices cover the definition of AI, i.e., agents that act in a human or rational way [5]. The first artifice, acting in a human way, invokes the Turing Test wherein human observers cannot distinguish whether the behaviour of a system was caused by either a human or a bot. The second artifice, of acting rationally, means that a rational bot can yield an optimum solution given a complex challenge and offering a wide range of corresponding solutions with varying degrees of risk.

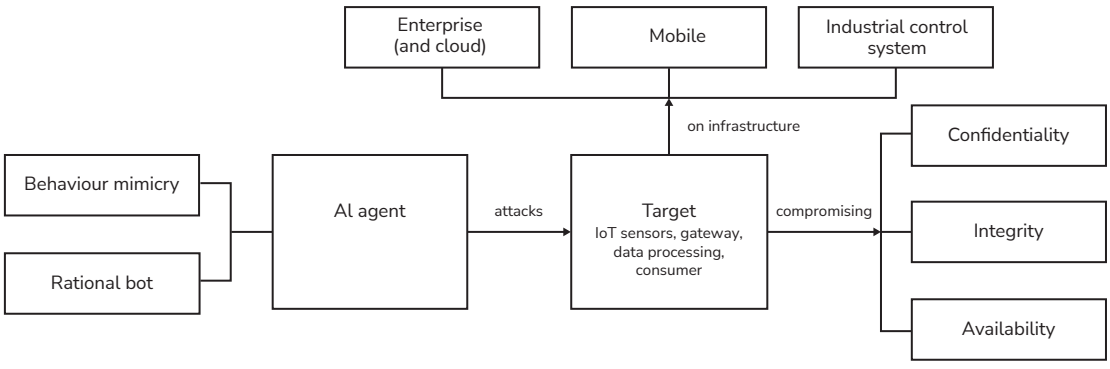


Figure 3. Techniques attacking the modern Internet.

A target operates in one of the three computing domain infrastructures, namely, enterprise (including Cloud), mobile, and industrial control

systems [6]. Fig. 3. shows that targets become victims when either one of the security goals is compromised. These are the confidentiality, integrity and availability of a computing system. Thus, an intelligent system becomes a victim when any one of the targets (i.e. IoT sensors, gateway, data processing, consumer) that is part of a computing infrastructure (i.e. enterprise/Cloud, mobile, industrial control systems) is attacked by some AI artifices (i.e. behavioural mimicry, rational bot), with the effect that one or more of the security goals (i.e. confidentiality, integrity, availability) is compromised. This means that attackers can employ a rational bot to advise on an optimum tactic flow out of many attack possibilities that have been described above. When engaging in a specific technique, the intelligent agents can deliberately find intrusion actions that produce data so as to get misclassified as normal. Thus, malicious AI agents capable of discovering the weakest link in a cyber system can be designed to launch adversarial AI attacks.

The work we present in this paper offers in-depth analysis of adversary approaches which exploit AI techniques to launch sophisticated cyberattacks. The approach allows us to see how the modern Internet phenomenon (Fig. 2.) and the complexity of the cyber kill chain [7] in penetrating cyber infrastructures can lead to the emergence of AI agents attacking the latter (Fig. 3.). We show that:

- as the modern Internet exchanges feedback data and knowledge, in addition to IoT-generated data, current devices are vulnerable to supply-chain compromise;
- the combination of possible tactics seeking to infiltrate cyber infrastructures is too complex to allow human analysts to understand zero-day attacks manually;
- in adversarial AI, agents can leverage the abundant data and the complexity of the problem domain.

1.1. Comparison with previous works

Discussions [8–10] to be found in the literature presented the Internet as a client-server architecture (Fig. 1.). Attacks were categorised on the basis of the control and modification of request packets [8]. From this point of view, adversarial AI techniques were tested only in respect of the meeting of data analytics goals, including as regards the efficacy of an intrusion-detection system and the ways in which network traffic can be misclassified [11, 12], or in relation to how a DoS attack is carried out through control of the volume of request payloads and their corresponding packet sizes [12, 13]. Adversarial AI was viewed in terms of its being a matter of finding data models to detect phishing or credit-card fraud [12, 14], rather than having an external IoT device manipulating the model. Malware was analysed using white-box approaches [11, 14, 15], rather than being seen from the point of view of a rational bot that can combine previously-known techniques from a knowledge base. Analyses of AI attacks on intrusion-detection systems were viewed from a one-sided

perspective [16] wherein datasets can be manipulated to produce attack data. This view alludes to the IoT perspective (as Fig. 2. shows) whereby response packets also act as knowledge to create adversarial data.

Other works [17–20] also fail to present background on how AI agents can leverage supply-chain compromise to target intelligent data-processing machines deployed on a range of infrastructures including autonomous systems and critical infrastructures. They have shown how adversarial AI agents can be used mostly for data-processing systems or for specified cyber infrastructures, as Tab. I. shows.

Table 1. Comparison with previous works.

Supply Chain	Supply Chain	Data Analysis
Other papers	[21]	[11–20]
Our contribution	Covered	Covered

Although the literature has discussed the circumstance that AI can be used for both attack and defence, few have examined the use in adversarial cyberattacks [21]. Adversarial AI techniques have allowed for the development of video games and natural language understanding [22], speech recognition, computer vision, online recommendation systems, and bioinformatics [23]. Most techniques that discussed cyberattacks observed from the data-analysis point of view (Tab. I.), within which AI models were challenged by reference to the ways in which request packets can be manipulated, rather than how external devices can infiltrate AI behaviours. In that circumstance, there was little discussion of ways in which AI has gained use in current adversarial cyberattacks, or indeed on methods developed to mitigate intelligent agents designed for such attacks.

2. Adversarial AI techniques used in Cyberattacks

Artificial Intelligence (AI) techniques range from mathematics and statistics to logic models, whereby procedures are encapsulated in an algorithm. The latter are known commonly as machine-learning algorithms, comprising both machine learning and AI interchangeably. Machine-learning algorithms analyse data in terms of samples and features. As an illustration, if a dataset is in the form of a table, the samples are the rows, and the features or dimensions are the columns.

We briefly introduce certain common AI techniques in the following paragraphs, before going on to discuss how they might be put to adversarial purposes.

Expert Systems represent one of the earliest computing techniques for decision-making. By way of a series of if-then-else flows, human experts are mimicked in reaching a final state, given a range of input data. In cybersecurity, such can serve as a knowledge base identifying asset vulnerabilities [24].

In turn, Particle Swarm Optimisation approaches [25] mimic the behaviour of social animals, in that each individual learns effectively from the others, with a view to optimum solutions being arrived at, e.g. as regards food. Such techniques were used for classification, weight optimisation, feature selection and dimensionality reduction [26].

Naïve Bayes [27] is a classic algorithm that gives acceptable results as data are classified. As such, it is used as a benchmark in comparing classification performance when a new machine-learning algorithm is developed.

Support Vector Machines [28] is a classification technique that can classify non-linear data. It is relevant to cybersecurity analysis because Internet traffic consists of heterogeneous data generated from a wide range of devices.

Artificial Neural Networks (ANNs) [29] analyse all features from each sample as a meshed network. This allows for flexible re-learning of new samples even after the model converges at the end of the training phase. This behaviour of ANNs is suitable for the analysis of Internet traffic in which there are rapid changes of the data pattern.

While Naïve Bayes and SVMs process the values directly from data features, ANNs can have multiple layers of intermediate features. Each layer can be designed to represent a set of features that is derived intuitively from the previous layer. Such layered networks are used in Deep Learning.

Deep Learning or Deep Neural Networks (DNNs) come with their derivatives, each with variations as to how networks are connected [30]. Their applications are discussed further in Section 2.1. Deep Learning networks have at least an intermediate, or a hidden, layer of units that is present between the input and output values. When the flow which adjusts the weight progresses in one direction from the input to the output layer, the network is called a feedforward network. If the adjacency of input values matters, then Convolutional Neural Networks (CNNs) become the architecture of choice. Instead of having a mesh of connections from the input values to the following layer as in DNNs, each unit in the CNN hidden layer is connected to a group of input values. As such, adjacent input values are captured as a spatial region. While CNN architectures represent spatial relationships, Recurrent Neural Networks (RNNs) consist of an architecture by which to model data with temporal characteristics. RNN outputs at time T are looped back, such that when unfolded its value can be calculated together with the input at time $T + 1$. On this basis, the network remembers and computes by reference to inputs from a series of time states. A derivative of RNN architecture is provided by Long Short-Term Memory (LSTM) networks, which can handle a longer chain of units without losing prior information. A Generative Adversarial Network (GAN) is a pair of ANNs wherein one network generates fake data samples that mimic the original training data, and the other network classifies the fake and original data. The two networks compete during the training iterations, with one network attempting to mislead the other. Thus, the generator aims to create fake samples that the discriminator cannot distinguish from the training data. The discriminator aims to classify the fake samples from this training data. As this section will discuss, GANs have attracted much attention in adversarial AI techniques.

2.1. Adversarial AI agents

Cyberattacks can leverage defensive AI techniques to compromise cyber systems. There are two characteristics of modern AI solutions that allow for the emergence of malicious AI agents, i.e. iterative learning and the use of a knowledgebase. Iterative learning allows devices to learn from the data generated from other ata-processing devices. As an example, defensive techniques such as anti-malware can be repurposed to develop new variants of mobile malware through iterative learning. In [31], the authors used Deep Learning to ascertain whether a malware variant was detected by anti-malware. The neural network iteratively mutated the variants by obfuscating their code until it was able to evade a group of anti-malware programs tested. Similarly, in [32], the authors used Genetic Programming to mutate executable programs. In this case, the subroutines of the programs constituted the chromosomes. They were selected and crossed over to create new malicious code, and then the resulting code was obfuscated. To test whether the code had become malicious, anti-malware was used twice, i.e., before and after the code was obfuscated. Iterating this process improved the selection of fitness values, in that a smaller number of malware detections was noted after the second test was conducted, as compared with the first one.

Inevitably, the use of a knowledgebase can also give malicious AI agents a competitive advantage. The authors of [33] described two types of real-life AI-based attacks that had occurred previously, i.e. attacks that took advantage of humans as the weakest link; and those that benefited from a rich knowledgebase as Fig. 4. illustrates. Using humans as the weakest link, in [33], attackers employed a tool to observe how a human user clicked and forwarded messages on social media. This allowed attackers to identify the most vulnerable target prone to a clicking-based phishing attack, and employed AI-based techniques to tailor highly relevant messages to the targets. In the second type, where the attack employed a knowledgebase, it was possible to describe a software vulnerability that had previously been proven to patch software. In a competition setting, the knowledgebase was employed to create autonomous attacks targeting and successfully compromising software systems belonging to other contestants. In [33], the authors also described a hypothetical case in which AI agents launch cyberattacks (Fig. 4.). It shows that worms, or codes that can spread autonomously to other systems, can automate the above cyberattack scenarios, e.g. by creating phishing emails or crippling target systems.

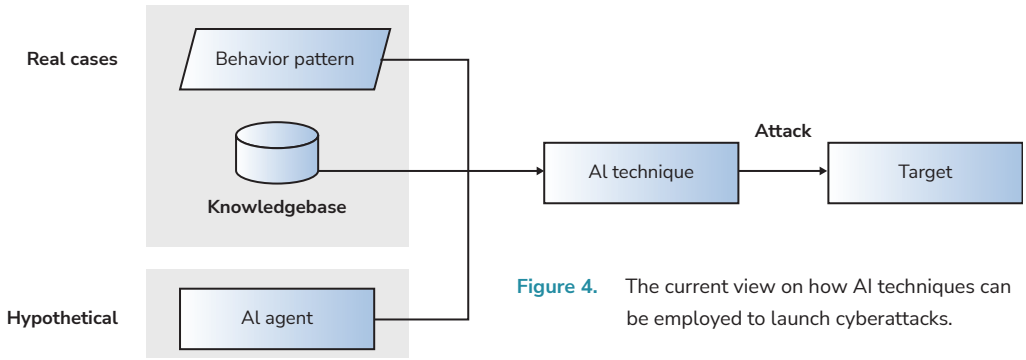


Figure 4. The current view on how AI techniques can be employed to launch cyberattacks.

Our study agrees with the above view that AI-assisted cyberattacks are increasingly a threat as they can conveniently circumvent existing security controls. Intelligent agents can engage in the autonomous targeting of weakest links in system, mimic legitimate behaviours, bypass intrusion-detection systems, and spread across different infrastructures. In this section, we show that current research has developed certain scenarios previously regarded as hypothetical. Fig. 5. illustrates the structure of the remaining discussion.

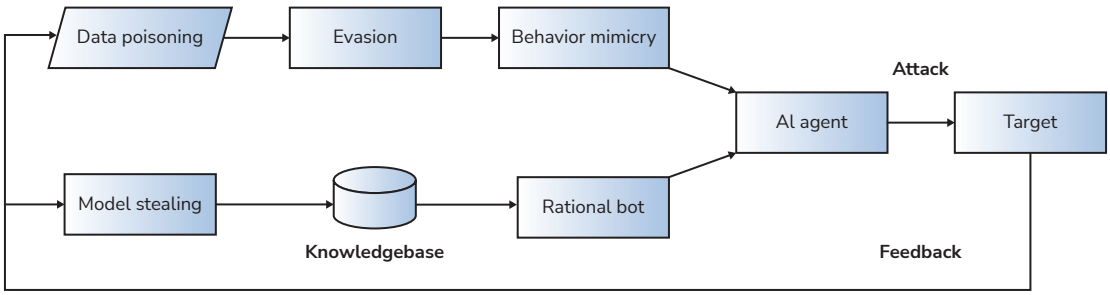


Figure 5 Intelligent agents launch cyberattacks through behaviour mimicry and rational bot techniques.

Fig. 5. shows that malicious AI agents launch cyberattacks through behaviour mimicry and rational bot techniques. As a result, the attacked target behaved differently in terms of computing output or performance. The AI agent captures these differences to optimise its attack strategy. The attack strategies are further applied to evasion, data poisoning, and model stealing techniques.

2.2. Evading detections by mimicking legitimate behaviours

Cyberattack detection has been described as detecting anomalous behaviour in networks or by users [34, 35]. Intelligent agents would mimic normal behaviour of networks, computer systems, or users, in order to bypass intrusion-detection systems. These agents are equipped with the statistical distribution of human-generated traffic patterns when online [34]. Intelligent machines would mimic the action of a human very closely [35]. Hence, the capability of an adversarial intelligent agent to evade detection can be attributed to behavioural mimicry.

The mimicking of legitimate behaviours can be made possible because the data generated by a device are no longer a mere response to a request packet as Fig. 1. shows. In today's Internet architecture (Fig. 2.), intelligent devices process request packets to generate insightful information (knowledge) in line with a device's data-processing behaviour. Thus, for example, a cloud service might send knowledge to a smartphone about which communication path would be optimum to traverse between two points.

The knowledge depends on the data-processing intelligence in observing the smartphone user's behaviour, e.g. as regards time of day, mode of transport, and most apt user preference between shortest path and journey time.

The data generated by a device thus depend on what that device has learned in addition to the request packet. It is therefore possible to train a device by mimicking some normal behaviour patterns in order to generate certain targeted data. This means it is possible to feed false data to the Cloud service (Fig. 2.) as generated from IoT sensors (Fig. 2.). Such a technique is known as the "supply-chain compromise" [36–38]. Behaviour mimicry is therefore an approach that can be taken in launching supply-chain compromises.

2.2.1. Data poisoning

The techniques to evade detection systems poison a system's input data. Poisoned data is basically contaminated data that can cause the detection system to misclassify inputs. Data poisoning assumes that the adversarial system has *a priori* knowledge of normal patterns. For example, in Internet traffic, normal network traffic patterns are those generated by human users as they browse websites. In [34], the authors demonstrated attacks that pre-empt a target's service successfully evading an intrusion-detection system even as the target was flooded with normal traffic, with the target caused to drop packets. Fig. 6. offers a statistical illustration of poisoned data. The black curve represents the distribution of a normal traffic feature value; the grey bar/curve represents attack values; and (left figure) the red threshold shows a detection system, which separates attack data from normal data. The right figure illustrates data poisoning. The normal values are contaminated with the attack values, evading the threshold bar, and rendering the attacks stealthy.

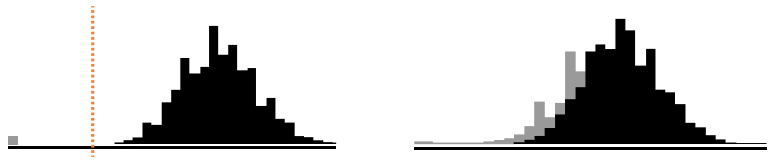


Figure 6. Left: a threshold (dotted vertical line) separates attack values (grey) from normal values (black). Right: attack values mimic the distribution of normal values. Adapted from [39].

2.2.2. Stealthy attacks

Fig. 6 summarises the method by which stealthy AI-based cyberattacks can be designed, i.e. through the supply of data whose anomalous value range overlaps with the acceptable range. While the above example [34] demonstrated that stealthy attack traffic can be prevented by the dropping of packets, the authors of [40] showed how dropped packets in

wireless networks can be made stealthy, causing the invariable blacklisting of legitimate nodes. In wireless networks, nodes (e.g. wireless devices) communicate within a certain communication range defined through by way of respective radio power values. Intrusion-detection systems can be deployed by means of collaborating nodes forming a series of overlapping radio range for scanning, which allows a node to cascade its network range of observation alongside neighbouring nodes. A malicious node can therefore be detected when it fails to forward packets within a time threshold to its neighbour. However, invoking set theory, the authors of [40] demonstrated that a malicious node located at an intersection of two sets of radio ranges can intentionally misroute packets such that they are forwarded to a victim node, causing the latter to drop packets and be blacklisted by the intrusion-detection system. This shows the way in which a malicious node that had learnt about the threshold value of a system and its position was able to affect the reputation of another node.

Stealthy attack methods are applicable to a wide range of cyber infrastructures, such as industrial control systems, facial recognition, and autonomous vehicles. In industrial control systems that monitor the degree of acidity (pH) of water, attacks can change the water pH values to a dangerous level [41], where the adversary is assumed to have compromised the pH meter device. The authors of [41] showed that a detection system that depends on a threshold value can be evaded by having the attacker adapt to the threshold value/range.

A case of the use of supply-chain attacks can be seen in the protection of cyber-infrastructures in which use is made of certain physical barriers, with a view to physical intruders being delayed, deterred and detected. Physical access, e.g. involving direct access to cyber-equipment, contributed to 56.3% of attack vectors in 2019 [42]. The mitigation of physical intrusions would entail installation of physical locks to deter and delay access, or cameras to detect presence. In this case, facial recognition can play a role in protecting cyber-infrastructures. Cameras can be programmed to recognise faces and raise alarms when they capture non-whitelisted facial images. However, in this case it is possible for cameras to be evaded to raise an alarm. Evading a classifier that recognises images of a person's face is to be done where an indiscernible image (such as a eye glass) is added to the input image [43], causing the classifier to recognise it as a different person. The authors of [43] first assumed the possession of the knowledge of the classifier, so that stealthy patterns might be designed. They used Deep Neural Networks to construct the classifier and to find a set of patterns, r, of the left-tail norm x, such that $x + r$ is classified into a desired class. Second, the study conducted a black-box test by applying the pattern (i.e. the eye glass) to a commercial face-recognition Cloud-based solution. Thus, the test relied on the software output as feedback to readjust the stealthy pattern r. However, the cloud classifier only outputs the top three classes, causing further difficulty with finding an indiscernibly modified input $x + r$. The study therefore used the Particle Swarm Optimisation algorithm that presented intermediate personified images, allowing each iteration to move away from the previous solution space and to approach the candidate solution more closely. As a result, 19 out of 20 images tested in the study proved to be evaded successfully.

Knowing the a threshold value of how much friction and wind can be tolerated by autonomous vehicles allows adversaries to fake a vehicle's positions without being detected. The authors of [44] demonstrated that adversaries can infer the total number of errors that can be tolerated by two target autonomous vehicles. The targets employed the Kalman filter, applied commonly in estimating the position of remote vehicles from the time lapse following on from the last collection of data. If an attacker compromises the control input/output of an autonomous vehicle, then the time-series data describing the vehicle's position can be derived, such that the error window size can be inferred. This allows the attacker to possess knowledge as to how much total error was tolerated within a time window. Injecting such stealthy errors into the controller can cause vehicles to deviate from their positions. As a result, a drone was, for example, shown to take a 50% longer time to accomplish a mission, while a rover took 30% longer [44]. In addition, the authors demonstrated that a large, poisoned dataset successfully evaded an intrusion detection system when it was subjected to drone memory, causing a drone to deviate 11 degrees when landing – a departure sufficient to prove fatal.

2.2.3. Perturbation

Another name for data poisoning is perturbation attack, as found in common use against industrial control-systems infrastructure. Perturbations are input noises whose range values are permitted by the system. In a network of water pipelines, the intentional perturbing of a water-meter reading can cause a machine-learning-based classifier to allow unusable water to be distributed to the population. A study [45] proposed that such a classifier should behave as a linear constraint to detect anomalies, given its mimicking of the water-flow behaviour in a pipeline network. While the water network in question was equipped with meter sensors by the provider, a linear constraint mandates that, for example, linear values are expressed as: *Meter1 > Meter2 + Meter3* is a legitimate pattern. Otherwise, the classifier detects an anomaly. The water network that the study observed was complex, consisting of 51 attributes, i.e., the values from 25 meters and the state of 26 actuators (i.e. valves that control flow), yielding a complex linear matrix. The study leveraged the system's tolerance for noise and normal fluctuations to perturb water-flow measurement, and ensure successful evasion where the detection of bad water flow was concerned. Similarly, perturbation of the voltage in an electricity network system can cause the system's classifier to misclassify electricity events identified in network traffic flowing into the grid [46]. In both cases, i.e. a network involving water [45] and electricity [46], the adversary can take an extra step in generating a normal pattern by assuming the classifier model. As is noted above, in the case of a water network, an adversary can assume a linear-constraint classifier, whereas in an electricity network, the adversary can assume a Convolutional Neural Network classifier, given that electrical-event behaviours are described ideally in time and space. Furthermore, the studies of both the water and the electricity network assumed the compromising of the systems' meters by an adversary, with

this allowing it to measure results obtained through system perturbation, and to generate poisoned data.

A variation for an adversary in modelling the target system is to assume feature ranking. In [47], the adversary perturbed the feature values describing events in an electricity network system, starting from the highest-ranked features. The perturbation of feature values could be achieved iteratively to the next ranked feature, until such time as misclassification of electricity events was observed, with this therefore aiding the adversarial objective of disruption/sabotage.

2.2.4. Fuzzing

The poisoning of input data can rely upon fuzzing techniques applied generally in software auditing. Fuzzing generates many input patterns that are input to a software program, so that many execution paths can be monitored for the purpose of bug-detection. The black-box fuzzing approach is dynamic analysis, by which poisoned inputs are fed into a program and the behaviour thereof is monitored, to confirm whether one of the security triads (i.e. confidentiality, integrity, and availability of the program) has been compromised. In [48], the authors describe an example in which a black-box fuzzing approach has been used to ensure that a program cannot be executed. Poisoned files are used as input to evade a program's parser recognising malicious files, with the result being for the program to execute the file. Test files were poisoned by fuzzing their bytes (i.e. random byte, delete, clone, or overwrite) from certain sections (e.g. the header). The poisoned files were still accepted for execution by the program, with the result that it was crashed successfully.

Table 2. AI techniques are used at all fuzzing stages.

Fuzzing stage	Challenge	Non-AI solutions [49]	AI solutions [50]
Seed file generation	Find the pattern that can save CPU time	Standard benchmarks; open-source samples	LSTM to learn known vulnerabilities in the sample
Testcase generation	Cover more program execution states	Dynamic taint analysis; probabilistic context-sensitive grammar	Neural networks to predict parts of greater vulnerability
Testcase filtering	Select inputs likely to find new paths	Hardware (Intel)	CNN to predict reachability of inputs
Operator selection	How to change input patterns efficiently	Generation-based: knowledge of program input is required	Mutation-based, derived from genetic algorithm
Exploitability analysis	Find vulnerabilities, not only crashes	Random-based fuzzing strategy	SVM and Bayesian to analyse features

AI techniques have played a significant role in exacerbating the difficulties noted previously in fuzzing steps. As Tab. II. shows, survey studies [49, 50] make it clear how each fuzzing step poses its own challenges and solutions. Fuzzing requires knowledge of how a target program is coded, and how it behaves under certain test patterns. As there are many cases to test, good initial patterns, or seed values, are required to efficiently find new execution paths to save computational resources (e.g. CPU time). Fuzzing is therefore an optimisation problem with multidimensional input vectors. Traditionally, certain assumptions have been made as to which input vectors can reveal program vulnerabilities efficiently; and some open-source seed patterns were adopted in consequence. The detection of new patterns capable of revealing vulnerability in software execution paths was achieved through random change of input vector values. In [50], the authors presented a survey reviewing 44 studies showing how AI techniques have the advantage of processing data as vectors, allowing many input patterns to be trained and labelled. Seed files can be represented as feature vectors, and good input patterns can be learned through training. Through the adoption of certain mutation-based algorithms, new feature vectors can be generated efficiently. These features gain analysis as AI, SVM, Bayesian and other strategies are used to select the fittest input values.

2.2.5. Discussion

We can make three observations from our analysis of the behaviour-mimicry techniques present in various IT infrastructures, namely that:

- normal patterns can be learned, whereas where behaviour mimicry is absent, they are either assumed, simulated or captured to create a dataset;
- data poisoning represents a subset of evasion attacks;
- the said mimicry of behaviour expands to mimicking machines.

Expanding on these points we first note how the adversary is in a position to learn. The distribution of normal patterns does not have to be assumed. Fig. 5. illustrates this, with the feedback arrow making this clear. In the modern Internet, target systems can be intelligent devices (e.g. smartphones) that send feedback information on receipt of a request message. Such feedback data are useful for the adversary to learn the threshold value of a detection system, and to eventually create a normal pattern dataset. For example, in autonomous vehicles, an attacker can infer the errors that the vehicle can tolerate by learning from its last location [44]. In industrial control systems, attackers can infer the tolerated noise in a water [45], or by observing water/electricity flows measured by the compromised meters in an electrical system [46]. Thus, data poisoning techniques can learn and create data, rather than merely assuming normal patterns.

Second, the goal of data poisoning in cybersecurity is the evasion of detection. This differs slightly from how attacks were defined outside the

cybersecurity context. From a data-analysis perspective, poisoning attacks take place at the training stage, while evasion attacks occur at the testing stage [17, 48]. This reflects the differing data-analysis strategies, i.e. white-box and black-box. As attackers in a white-box setting have knowledge of the AI model training stages, training data can be poisoned, and inputs classifiable as false negatives can then be supplied at the testing stage. On the other hand, cyberattacks should be viewed from the black-box perspective. As the authors of [51] discussed, attackers in a black-box setting can estimate the allowed data values for poisoning, by learning from the feedback data. This agrees with the first observation mentioned above, that normal patterns can be learned/estimated.

Third, AI-assisted cyberattacks are concerned with mimicking, not only human behaviour, but also machine behaviour. Techniques applied in evading intrusion-detections mimic, not only human browsing behaviour [34], but also machine behaviour involved in enterprise network systems [40], industrial control systems [41], and autonomous systems [44]. The definition of intelligent systems may therefore need expanding, to include systems that can convince other machines of their status as legitimate peers.

2.3. Rational bots

In the modern Internet architecture, AI-assisted black-hat agents enjoy an incomparably greater advantage over white-hat human analysts. Today's Internet creates a network of networks too complex for manual traffic data analysis to be performed. Traffic analysis has become even more complex because devices are interconnected with others across various heterogeneous infrastructures. Illustrating this, in [52], the authors proposed a system that detects whether an elderly person has fallen accidentally. This system consists of a wearable device (mobile) connected to a digital gateway at home with a view to (remote) healthcare of the elderly being facilitated. The gateway sends data to a Cloud service that determines whether the device user has fallen accidentally at home, or merely taken a rest-motivated lie-down in a deliberate and intentional way. When a true fall is detected, the Cloud service sends an alert message to family members and to an emergency centre (critical infrastructure). This illustration shows that complex and diverse IT infrastructures are involved in providing a timely, critical solution. White-hat human analysts can become overwhelmed when attempting to find patterns from intrusion logs across different infrastructures and platforms. As such, they have built knowledge bases to model cyberattacks [53]. Such bases allow cybersecurity analysts to share their knowledge about how new techniques and tactics have succeeded in compromising a target, and allow them to engage in incident handling. However, such knowledge bases can also be leveraged by black-hat AI agents, who would like to build rational bots capable of discovering security holes in a complex system whilst focusing on the weakest link. Such a tactic for discovering security holes that are outside the current compromised system is known as "lateral movement" [54, 55, 56]. AI agents can be used by adversaries to engage in rational detection of a system's weakest link, with lateral-movement attacks then performed.

2.3.1. Model stealing

Model stealing is a technique by which to create a model mimicking algorithms adopted for detecting attacks, as Fig. 7. shows. A defensive classifier for detecting attacks can classify its input data to attack/legitimate as output. Both the input and output data-vectors create a new dataset for an attacker. The output data acts as the label to each input, such that the new dataset is a labeled one capable of being used to train AI techniques. The attacker can use this new dataset to train an adversarial model mimicking a legitimate intrusion-detection system. Deep Neural Network techniques are adopted as an adversarial model to mimic a machine learning-based classifier. In [57], the authors used a classifier for machine learning-based intrusion-detection, and the proposed stealing model yielded results of 99.59% accuracy.

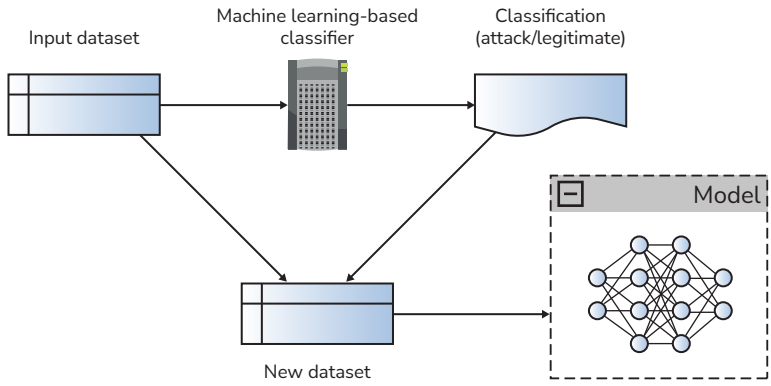


Figure 7. Model stealing, adapted from [45].

As Tab. III. makes clear, most challenges in model stealing appertain to knowledge of input data accepted by the victim service e.g. a machine learning-based classifier, so that the adversary knows what to query. On one hand, exact input data is required for the classifier to be stolen; but on the other there is a need for knowledge on classifier, in order for acceptable input data to be produced. Construction of that input data entails estimation of its range of parameters in a multidimensional space, as well as limitation on the amount of required data samples with a view to computational complexity being minimised. Input-data generation is one of the issues faced by AI systems. To address this, the author of [58] adopted a smaller external dataset to estimate the input range of a dataset used by a victim service hosted in the Cloud. Generative Adversarial Networks (GANs) were deployed to estimate parameters in the stolen model, with the external dataset as the input. The adversary had no *a priori* knowledge of how much of the external data overlapped with the input data in the Cloud. To overcome this, a knowledge distillation technique [59] was adopted to measure the loss, i.e. the difference between the output from the GAN model and the output from the Cloud model. Smaller differences denote correct estimation of output values. Then, the output data that the victim was able to label correctly label received a higher weighting than other

output data. This allowed GANs to estimate the parameters in the stolen model, thereby generating synthetic input data. In return, the latter yielded a better stolen model. The quality of the input data was measured by reference to the inception score-higher numbers come from a lower entropy value when the joint probability of the synthetic and victim data is high. The work showed that synthetic data achieved 60.58% higher inception-score values than the external data.

Table 3. Model stealing issues depend on where the data is processed.

Location of data transformation activity	Challenge	Solution
Edge	Synthesising input Data	Minimising [58] or maximising [60] the entropy of joint probability distribution of victim-synthetic
Cloud	Minimising the amount of traffic for query	Finding a subset of input data [62]
IoT (MOBILE)	Noisy media, incomplete classification data	Reconstructing the data employing human judgment in viewing images [63]

Similarly, the authors of [60] employed an external dataset and the knowledge-distillation technique to create synthetic data. In contrast to the work of [58], this seeks to maximise the difference between the victim's output data values and the output values obtained from the stolen model to measure loss. The measure employed zeroth-order gradient estimation [61] to update the weight that maximises the entropy of probability distribution between the victim and stolen model output. By training both models an input dataset was generated which was then used in training the stolen model. The synthetic dataset maximises the learning of the stolen model, and therefore creates a highly accurate stolen model.

When the machine learning-based classifier (Fig. 7) is located in the Cloud, the challenge of stealing the model is to query the Cloud service discretely. Many queries to the cloud can trigger an alarm. In [63], a discrete querying of the model was developed by employing the transfer-learning technique, which finds a subset of the input data such that the total number of samples is minimised. With DNN, input data was pre-trained to yield a new dataset of lower dimension. The authors of [62] demonstrated that the stolen model of a Cloud-based image classification model achieved accuracies of 83.73%. The number of queries was 1290, compared with 5000 for other comparable results, which suggests that the model was stolen discreetly.

When the machine learning-based classifier is located on an IoT platform, the challenge of model stealing involves handling of incomplete data as obtained from a classifier's output. Wireless data is noisy, resulting in the capture of incomplete data properties, as [63] demonstrates.

Thus, when there is only a small amount of classification data available, generation of the stolen model can be aided by human judgement. In [63], the authors reconstructed a model by which to predict lung cancer from pulmonary data. The captured data was displayed as 3D images, allowing humans to add certain properties, such as marking of lesions, with this ensuring the creation of labelled input data. With this data, the authors developed a stolen model, employing a Convolutional Neural Network. The technique showed that the stolen and original model differed by 0.3% in terms of accuracy.

Tab. III. summarises the discussion on model stealing. The challenges here can be viewed from the location of the machine learning-based classifier that transforms data, i.e. either at the edge, in the Cloud, or on a mobile device. The issue is based fundamentally upon technique for data acquisition. The more remote the classifier is, the more limited the data acquired. Thus, AI techniques play a crucial role in estimating the data parameters to help an adversary achieve its goals.

2.3.2. Discussion

The above discussion shows that model stealing techniques collect, estimate and create data. It is therefore possible for data to be collected into a rich database, oracle or knowledgebase (as the authors of [33] demonstrated), to provide data that facilitates future model stealing-based adversarial activities. As Fig. 5. illustrates, when such a knowledgebase also incorporates solutions from the evasion techniques, an adversarial bot or malware can act rationally with the aim of carrying out cyberattacks. If the rational bots or malware knows the victim's model, it can select the best tool for reconnaissance, discretely open ports and replicate itself, maintain persistence on the target system, escalate privilege, and conduct lateral movement to attack other platforms. An AI agent that is equipped with both evasion and model-stealing techniques can mimic the normal data parameters when pursuing the above attack chain, creating undetectable attacks.

2.4. Solutions to mitigate adversarial AI attacks

While the previous subsection formulates AI-based cyberattacks, this subsection discusses their mitigation techniques to defend against adversarial AI attacks. Principally, these methods derive from, first, the assumption that the victims had significantly more knowledge about their own network/system parameters than their adversaries. Second, that the adversarial techniques would have some disadvantages inherent to them. Thus, the mitigation techniques leverage these adversarial techniques' disadvantages.

2.4.1. Feature definition

A set of features is what enables a machine-learning technique to engage in classification. When an adversary has obtained the feature set, they can

mimic the way the machine learning technique classifies data. In this case, the defenders' option would be to redefine the set of features for data analysis. This is made possible when, for example, a new technology is introduced, causing the defender to re-analyse data. In [34], the authors showed that, when the new web communications protocol HTTP/2 was introduced, the traffic pattern was different from its predecessor, HTTP/1.1. This allowed adversaries to create attack traffic undetectable by machine-learning techniques. Thus, the authors of [34] proposed a new set of features to allow for the detection of the stealthy attack traffic.

The defining of a new set of features follows the data mining technique, such as the one described in [64]. A feature is identified either by observation or intuition. A good feature has a value distribution that can identify the intended class (e.g. attack or legitimate) closely. For example, a feature with wide data distribution can allow a threshold to be placed for classification (Fig. 6). When data is multidimensional, a set of identified features is ranked according to how well the combination can lead to a classification. Algorithms such as information gain [65] are used commonly in ranking features.

Both the adversary and the defender can find a set of features for attack modelling as well as for mitigation, as in the case of Cloud services. Cloud services see incoming requests from any client connecting to them, of either legitimate or fake status. When a machine-learning service is deployed in the Cloud, it is vulnerable to fake queries that can be crafted for the purpose of model stealing. That is, from the adversary perspective, the distribution of the input dataset can be inferred (as discussed in Section 2.3), allowing adversaries to define morphed features. To detect adversarial query traffic from the defender's perspective, the authors in [62] analysed the distribution difference between legitimate and adversarial query traffic. They then proposed a set of features by which to detect the adversarial query traffic, allowing them to detect abnormalities in the input data with 92% accuracies.

2.4.2. Monitoring

As discussed in [66], network management involves configuration and measurement. Monitoring is the activity of collecting and analysing network measurements, which depict the network's behaviour and performance. Monitoring involves the collection, analysis, and presentation of data (Fig. 8). The collection layer captures network traffic; the analysis layer extracts network traffic and converts the same into relevant data; and the presentation layer provides for meaningful (e.g. graphic) representation of the data that interprets network behaviour and performance. Because such network monitoring is sourced from network traffic generated from the whole network under management, cyber-network owners have better knowledge of their own network than the adversary.

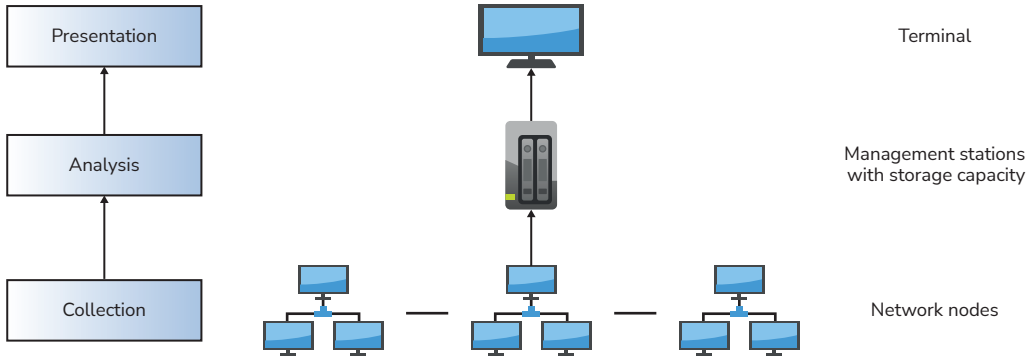


Figure 8. Monitoring as the management layer in measuring network performance, adapted from [66].

In [40], the authors observed that the adversaries can only access a subset of adversarial nodes situated in the network to generate stealthy attack traffic. A compromised node can manipulate only a subset of its neighbouring nodes to believe (i.e. measure) that certain packets have been forwarded. However, not all nodes can be manipulated in a single broadcast, because traffic-volume configuration and historical measurement differ from one node to another. Thus, the authors proposed to expand the monitoring of the network so that more neighbouring nodes count the number of packets forwarded from each node. If one node has differing belief than another as to the number of packets being forwarded by a monitored node, then packet dropping had occurred.

Similarly, in critical infrastructure networks, the authors of [41] observed that attackers face non-negligible risks of being detected if they generate attack traffic imprudently. This is because the behaviour of the network can become unpredictable to the attacker. In simulating attacks, the author of [41] assumed that the attackers know the distribution of the input sensor values (with consumer behaviour considered to consume utility over time). However, sensor values depend on complex interactions with other sensors in real network settings. Attackers have less knowledge than defenders in predicting consumer behaviour and random perturbations in the network. Thus, the deploying of an extensive monitoring system throughout the network can confer advantage upon the defender.

2.4.3. External validation

External validation offers an innate defence against adversarial AI, because the efficacy of research findings may not be as valid when applied to real settings. Research conducted by the authors of [41], for example, simulated the critical infrastructure out of a room-sized lab, showing how pH levels in water can be manipulated. The infrastructure consisted of 6 main Programmable Logic Controllers (PLCs), one of which was to control the pH level. The study [41] assumed that a man-in-the-middle attack succeeded in controlling the PLC for pH level, causing the water acidity to change linearly, such that $pH_{time+1} = pH_{time} + delay$. Such lab-based behaviour is

more predictable than that in real settings, where there would be more PLCs, constraints placed on how long a PLC can be turned on/off, and complex filtering systems for neutralizing water parameters. This would all ensure the potential failure of an attack under real settings.

In [60], the authors recognised that the solutions in model stealing are impractical in reconstructing real-world image classifications. The problems are that the input dataset (Fig. 7) is not readily available to the attackers. Attackers only assume partial availability of the data, or the availability of a similar dataset. As such, attacks become ineffective when deployed in real settings. Furthermore, the authors in [60] discussed the fact that, while GANs were mostly used to construct a stolen model, this is not ideally applied to real image classification settings because the dimensionality of the generator's parameters can be in the order of millions.

Recall from Section 2.2.2. that image recognition can play roles in protecting cyber-infrastructure. In [43], a study showed how to evade a machine-learning technique that is recognising a person's face. The study used data from images taken with room lighting without exterior windows to prevent extreme lighting variations. The persons whose faces were taken as data samples must maintain a neutral expression and were stood at a predetermined distance from the camera. Such data are not valid externally; the authors acknowledged that detection using real outdoor images is challenging. This means that one defensive method against adversarial AI attacks is to embrace the degrading external validity, for example by creating a more complex system.

2.4.4. Alteration of parameters

As Fig. 6 shows, adversaries can infer a classifier's parameters, such as the threshold, to mimic the distribution of target data. The approach to defending against such a scenario is thus to change parameters, either as the combination of parameters used by the classifier or as parameter values. In [44], the adversaries inferred both the threshold and the window size of the target data, allowing them to inject false data to change a drone's position. However, when these parameters were changed as part of a defensive method, the room for manoeuvre to inject false data was decreased. This is illustrated in Fig. 9. Under a normal situation, a drone's position does not deviate from the allowed position threshold. When false data were injected, however, the position deviated but was still below the threshold (undetected attack 1, Fig. 9.). Even though a drone's position deviated above the threshold, this did not raise an alarm because the event was construed as just a fraction of the time window (undetected attack 2, Fig. 9.). To mitigate these attacks, the authors of [44] suggested having an adaptive threshold and variable-size window which will result in the impact of the attack being diminished. In an adaptive threshold (mitigation 1, Fig. 9.), the allowed position deviation threshold was changed, causing the false data to fall above it. In a variable-size window (mitigation 2, Fig. 9.), the injected false data still reached above the threshold at all times, precisely because the window size was reduced.

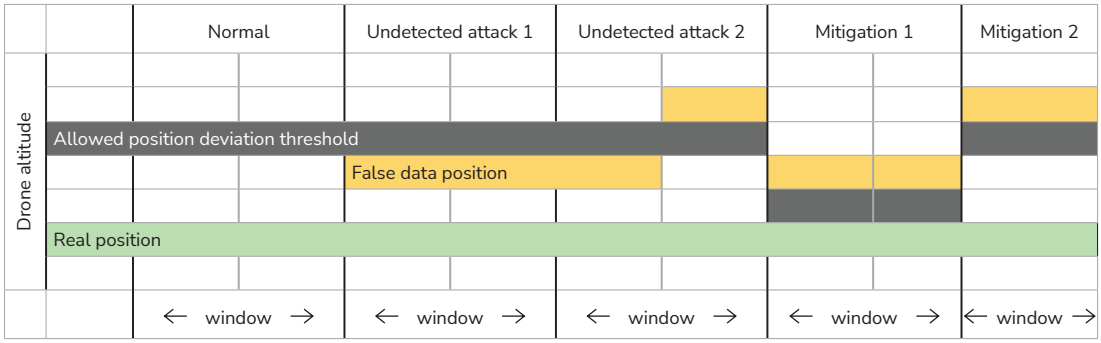


Figure 9. To detect anomalies a defensive method can alter either the threshold or the window size.

The second method is to change the parameter values of the machine-learning model relevant for classification prediction to a higher level of precision. As the authors of [58] have shown, Cloud machine learning services round their values to two decimals to provide only the necessary information. Yet, Cloud services can make internal use of the higher-precision values to generate their output data. By supplying only a fraction of the data at the output, i.e., rounded values to a lower precision, Cloud services can defend against model-stealing attacks.

2.4.5. Adversarial training

In adversarial training [67], adversarial attack techniques are used to train the system to be defended. The technique uses perturbed input data, which represent the attack samples, to cause machine-learning models to misclassify. Perturbing input data is crafting data samples such that their feature values are modified by a small deviation from their original value, causing the machine-learning algorithm to create a deviated / wrong function, to be used as the model. Collectively, the deviated values maximise the loss between the intended function and the modified function. Adversarial training is to generate a sizeable, perturbed input dataset, label the correct class, and use such perturbed input data as the training data to train the machine-learning model. Thus, the system becomes more secure because attack samples have been seen during the training phase. The technique is considered the accepted procedure by which to defend against the perturbation attacks discussed in the previous section. In this case, the system has seen perturbed inputs.

Traditionally, adversarial training was used for image classification in computer vision [67]. The technique is now adopted in sensitive domains such as cyber-physical systems [45], critical infrastructures [46], and industrial-control systems [47]. With adversarial training, the training data can be perturbed to represent the complexity of the physical system, e.g. in relation to actuator constraints in water systems [45]; voltage, current data, short-circuit fault, line maintenance, remote tripping, and relay settings in a power grid [46, 47].

2.4.6. Patching software

A patching of software vulnerability represents an effective approach by which to defend against fuzzing attack. Conceptually, this is like the adversarial training described above; with software being fuzzed to find, and eventually patch, vulnerability. There are two advantages to the patching of software. First, non-limitation solely to the securing of vulnerability, but also an increasing of complexity in the software execution path. The disadvantage of fuzzing is that it relies on code coverage [49]. This is to say that, the larger the code, the less successful is the fuzzing attack. Thus, fuzzing would not discover all the execution paths of the software following software patching. Second, the authors of [50] point out that the application of different AI techniques in fuzzing the same problem can lead to significant differences in the discovery of execution paths for attacks. An increases in the number of execution paths as software is patched can lead to a lowering of the success rate where fuzzing attacks are concerned.

2.4.7. Defensive distillation

Rather than having a machine-learning model that outputs a class with a high probability, the defensive distillation technique [59] suggests that the model smooths the probability of the output class. This causes the probability of the model generating one class to be similar to the probability that the model generates the other classes. The technique is thus suitable when it comes to defending against model-stealing attacks. One variation of the technique [58] is for the classifier to output only the top-n classes with the highest probabilities. This would limit the adversary's knowledge of what classes would have a low probability with a given sample. The technique can be enhanced further where only the top-1 class is outputted. Although in [58] the authors observe that there was no significant advantage in defending against model stealing attacks using the top-1 defence, they still believe this to be the logical defence technique, given the way it provides users with very limited information.

3. Future directions

The concept of AI-based cyberattacks has emerged from the convergence of AI algorithms, a rapid increase in computational power, application development and operation advancements, and the ready availability of AI-based implementations for adversarial adoption.

Future AI-based attacks are also determined by attacker motives. An increasingly significant threat is posed by lucrative opportunities for the adversary through cyber threats posed in contemporary times that involve hijacking of systems and encryption of user data, even as the latter are held for ransom (i.e. through ransomware), with user payments in cryptocurrency form demanded. Such opportunities may emerge through AI-based attacks that assess the vulnerabilities of victim machines, in advance of their sending a ransomware payload through to them. Motives other than financial gain

can include terrorism, business competition (e.g. as bots spread fake news), hacktivism or the expression of political views. The adversarial AI techniques discussed in this work (Section 2.1.) can be used to achieve such motives.

As we have discussed through the analysis reported in this contribution, the ability of computing platforms to thwart the AI-based cyberattack spectrum depends on the following observed aspects:

1. the design of computing platforms resilient to AI-based adversarial threats, not as an afterthought to production, but rather via an awareness that everybody has responsibility (as part of the DevSecOps paradigm);

2. the design of applications (web-based, mobile, and Cloud) that are resilient to AI-based cyber-threats, by way of the prevention of data capturing and fostering of attacker learning through the provisioning of feedback, i.e. a reduced amount of feedback data provisioned to end-users given the possibility of comprising both legitimate and adversary class;

3. the design of network security controls adopted in a network, with a view to cyber-threats arising through AI engine exploitation being thwarted (future directions for such activity may include egress and ingress packet-filtering based on detection of anomalous feedback patterns (statistical as well as pattern-based) that are moving through the network;

4. the identification of opportunities to obfuscate-neural network operations, parameters, and generated outputs, and the adoption of a black box-based framework by which to prevent the adversary from exploiting system weakness in provisioning of clearly indicative data to the adversary;

5. adoption as common practice of security by design, even as heterogeneity is incorporated in the nature and type of IT and Operations Technology (OT) devices that comprise a modern-day Information Communication Technology (ICT) platform (i.e. IoT and edge devices, digital controllers, Supervisory Control and Data Acquisition (SCADA) systems, Cloud servers and mobile devices) – design should include options to prevent AI-based cyber threats from being perpetrated against the holistic platform identified above.

Future adversarial AI attacks will become more pervasive over time. In modern Internet settings, the generation, exchange and processing of data rely on remote data operations, including those found in the evolving discipline of Industrial IoT (IIoT). As we have discussed in Section 1, each data-exchange end point on the Internet is vulnerable to exploitation. Modern society has become more dependent on the integrity and availability of such cyber-services as are seen in banking and critical infrastructures. For example, the integration of IoT devices with back-end Clouds is a common practice in contemporary computing domains including critical infrastructures, as where the aims are electricity distribution, load

balancing, and feed-in tariff in smart grids. The ability of the adversary to determine IT/OT vulnerabilities within a smart grid may prove catastrophic to the routine operations of critical infrastructure, which is essential to provision routine services such as electricity distribution to citizens.

Another emerging field of study is the digital forensic readiness of AI-based systems. Essentially, the vulnerabilities of such systems can be exploited by the adversary through the adoption of technologically-advanced tactics, including the circumvention of facial recognition systems, the bypassing of the security controls of AI systems, and the deliberate injection of falsified data into the communication stream. By analysing such empirical data, it is possible to hoard the right data types and data artefacts as may help a digital forensic investigation undertaken as part of post-incident analysis.

4. Conclusion

AI techniques have gained use, not only to defend traditional network systems, but also to attack their implementations. This is made possible because the modern Internet is exchanging not only raw data, but also processed data such as that generated by Cloud-based machine-learning services. This phenomenon is seen to affect the cyber infrastructure comprising enterprise, mobile, and autonomous systems, as these engage in the exchange of both sensory data and AI analytical data. Such infrastructure has become the playground for AI-based cyberattacks. The number of possibilities to attack – from reconnaissance, execution, persistence, privilege escalation, command/control, to data exfiltration-manifests as too large a spectrum for humans to analyse zero-day attacks. However, adversaries are one step ahead, using knowledgebase and known AI techniques to launch AI-based cyberattacks.

AI techniques are suitable for defining attack vectors because they can handle large volumes of data. In this paper, we described machine-learning-based techniques for adversarial AI. These techniques are adopted by the adversary to carry out adversarial AI attacks, with the derivatives of neural networks playing a significant role in fostering the development of novel AI attack vectors incorporating both spatial and temporal data in the emulation of legitimate data. We further classified Adversarial AI into behaviour mimicry (which employs stealthy attacks, perturbation and fuzzing techniques) and rational bot (which employs the model-stealing technique). Behaviour-mimicry techniques aim to resemble normal data, thus infiltrating victim machines. Rational bots employ the knowledgebase that was obtained from the model-stealing technique to facilitate the design of zero-day attack vectors.

Also discussed here are 7 methods by which to defend against AI-based cyberattacks. The mitigation approaches leverage the disadvantages understood from the adversarial AI techniques. And, even in the face of these countermeasures, adversarial AI attacks will be – as we indicate – more pervasive, as society becomes more dependent on cyber-data exchanges that offer a plethora of opportunities for adversaries to further their motives effectively.

Acknowledgments

We thank the anonymous reviewers for their valuable comments of assistance to us in improving the content, organisation, and presentation of this paper. Sherali Zeadally was supported by a Fulbright U.S. Scholar Grant Award administered by the U.S. Department of State's Bureau of Educational and Cultural Affairs, and through its cooperating agency the Institute of International Education ("IIE").

References

[1] I. Novikov. (2018). *How AI Can Be Applied To Cyberattacks* [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2018/03/22/how-ai-can-be-applied-to-cyberattacks/#3211152d49e3. [Accessed: Sep. 28, 2022].

[2] S. Zeadally, E. Adi, Z. Baig, I. A. Khan, "Harnessing artificial intelligence capabilities to improve cybersecurity," IEEE Access, vol. 8, pp. 23817–23837, 2020, doi: 10.1109/ACCESS.2020.2968045.

[3] J. Hobbs. (2018). *AI Enters the Cyber Attack Realm* [Online]. Available: https://www.afcea.org/content/ai-enters-cyber-attack-realm. [Accessed: Sep. 28, 2022].

[4] E. Adi, A. Anwar, Z. Baig, S. Zeadally, "Machine Learning and Data Analytics for the IoT," *Neural Computing & Application*, vol. 32, pp. 16205–16233, 2020, doi: 10.1007/s00521-020-04874-y.

[5] S. Russell, P. Norvig, *Artificial intelligence: a modern approach*, no. 4London: Pearson Education, 2020.

[6] *MITRE. MITRE ATT&CK* [Online]. Available: https://attack.mitre.org/matrices/. [Accessed: Sep. 28, 2022].

[7] *Lockheed Martin, The Cyber Kill Chain* [Online]. Available: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html. [Accessed: Sep. 28, 2022].

[8] I. Corona, G. Giacinto, F. Roli, "Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues," *Information Sciences*, vol. 239, pp. 201–225, 2013, doi: 10.1016/j.ins.2013.03.022.

[9] M. Babar, M. Sohail Khan, "ScalEdge: A framework for scalable edge computing in internet of things–based smart systems," *International Journal of Distributed Sensor Networks*, vol. 17, no. 7, pp. 1–11, 2021, doi: 10.1177/155014772110353.

[10] J. Neeli, S. Patil, "Insight to security paradigm, research trend & statistics in internet of things (iot)," Global Transitions Proceedings, vol. 2, no. 1, pp. 84–90, 2021, doi: 10.1016/j.gltp.2021.01.012.

[11] I. Rosenberg, A. Shabtai, Y. Elovici, L. Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–36, 2021, doi: 10.1145/3453158.

[12] A. McCarthy, E. Ghadafi, P. Andriotis, P. Legg, "Functionality-preserving adversarial machine learning for robust classification in cybersecurity and intrusion detection domains: A survey," *Journal of Cybersecurity and Privacy*, vol. 2, no.1, pp. 154–190, 2022, doi: 10.3390/jcp2010010.

[13] E. Alhajjar, P. Maxwell, N. Bastian, "Adversarial machine learning in network intrusion detection systems," *Expert Systems with Applications*, vol. 186, pp. 1–25, 2021, doi: 10.48550/arXiv.2004.11898.

[14] H. Navidan, P. F. Moshiri, M. Nabati, R. Shahbazian, S. A. Ghorashi, "Generative adversarial networks (gans) in networking: A comprehensive survey & evaluation," *Computer Networks*, vol. 194, no. 3, pp. 1–26, 2021, doi: 10.1016/j.comne vol.2021.108149.

[15] D. Li, Q. Li, Y. Ye, S. Xu, "Arms race in adversarial malware detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–35, 2021, doi: 10.1145/3484491.

[16] M. Pawlicki, M. Choraś, R. Kozik, "Defending network intrusion detection systems against adversarial evasion attacks," *Future Generation Computer Systems*, vol. 110, pp. 148–154, 2020, doi: 10.1016/j.future.2020.04.013.

[17] Z. Guan, L. Bian, VOL. Shang, J. Liu, "When Machine Learning meets Security Issues: A survey," 2018 International Conference on Intelligence and Safety for Robotics, Shenyang, 2018, pp. 158–165 [Online]. Available: https://ieeexplore.ieee.org/document/8535799. [Accessed: Sep. 28, 2022].

[18] G. Apruzzese, M. Colajanni, L. Ferretti, M. Marchetti. (2019). "Addressing Adversarial Attacks Against Security Systems Based on Machine Learning," 11th International Conference on Cyber Conflict (CyCon), pp. 1–18 [Online]. Available: https://ccdcoe.org/uploads/2019/06/Art_21_Addressing-Adversalial-Attacks.pdf. [Accessed: Sep. 28, 2022].

[19] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, J. D. Tygar, "Adversarial machine learning," In Proceedings of the 4th ACM workshop on Security and artificial intelligence, 2011, pp. 43–58.

[20] V. Duddu, "A Survey of Adversarial Machine Learning in Cyber Warfare," *Defence Science Journal*, vol. 68, no. 4, pp. 356–366, 2018, doi: 10.14429/dsj.68.12371.

[21] A. Szychter, H. Ameur, A. Kung, D. Hervé. (2018). *The Impact of Artificial Intelligence on Security: a Dual Perspective* [Online]. Available: https://www.cesar-conference.org/wp-content/uploads/2018/11/articles/C&ESAR_2018_J1-03_A-SZYCHTER_Dual_perspective_AI_in_Cybersecurity.pdf. [Accessed: Sep. 28, 2022].

[22] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436–444, 2015, doi: 10.1038/nature14539.

[23] I. Goodfellow, Y. Bengio, A. Courville, *Deep learning*, *MIT* press, 2016.

[24] M. M. Gamal, B. Hasan, A. F. Hegazy, "A security analysis framework powered by an expert system," *International Journal of Computer Science and Security (IJCSS)*, vol. 4, no. 6, pp. 505–527, 2011.

[25] J. Kennedy, R. Eberhart. (1995). "Particle swarm optimization," *Proceedings of ICNN'95 – International Conference on Neural Networks*, vol. 4, pp. 1942–1948 [Online]. Available: https://ieeexplore.ieee.org/document/488968. [Accessed: Sep. 28, 2022].

[26] M. H. Nasir, S. A. Khan, M. M. Khan, M. Fatima, "Swarm intelligence inspired intrusion detection systems—a systematic literature review," *Computer Networks: The International Journal of Computer and Telecommunications Networking*, vol. 205, pp. 108708, 2022, doi: 10.1016/j.comnet.2021.108708.

[27] VOL. Bayes, LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, In a letter to John Canton, AMFR S, Philosophical transactions of the Royal Society of London, vol. 53, pp. 370–418, 1763, doi: 10.1098/rstl.1763.0053.

[28] C. Cortes, V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[29] W. S. McCulloch, W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, vol. 4, pp. 115–133, 1943, doi: 10.1007/bf02478259.

[30] I. H. Sarker, "Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective," SN Computer Science, vol. 2, no. 3, pp. 1–16, 2021, doi: 10.1007/s42979-021-00535-6.

[31] Z. Fang, J. Wang, B. Li, S. Wu, Y. Zhou, et.al., "Evading Anti-Malware Engines with Deep Reinforcement Learning," *IEEE Access 7*, 2019, pp. 48867–48879, 2019. doi: 10.1109/ACCESS.2019.2908033.24.

[32] S. Sen, E. Aydogan, A. I. Aysan, "Coevolution of Mobile Malware and Anti-Malware," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2563–2574, 2018, doi: 10.1109/TIFS.2018.2824250.

[33] E. Zouave, M. Bruce, K. Colde, M. Jaitner, I. Rodhe, "Artificially intelligent cyberattacks", *Stockholm: Totalförsvarets forskningsinstitut FOI* [Online] Available: https://whttps://www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf [Accessed: Sep.28, 2022].

[34] E. Adi, Z. Baig, P. Hingston, "Stealthy Denial of Service (DoS) attack modelling and detection for HTTP/2 services," *Journal of Network and Computer Applications*, vol. 91, pp. 1–13, 2017, doi: 10.1016/j.jnca.2017.04.015.

[35] A. M. Turing, "Computing machinery and intelligence," in *Parsing the turing test*, doi: 10.1007/978-1-4020-6710-5_3.

[36] MITRE. *Supply Chain Compromise for Enterprise* [Online]. Available: https://attack.mitre.org/techniques/T1195/. [Accessed: Sep. 28, 2022].

[37] MITRE. *Supply Chain Compromise for Mobile* [Online]. Available: https://attack.mitre.org/techniques/T1474/. [Accessed: Sep. 28, 2022].

[38] MITRE. *Supply Chain Compromise for Industrial Control System* [Online]. Available: https://collaborate.mitre.org/attackics/index.php/Technique/T0862. [Accessed: Sep. 28, 2022].

[39] E. Fix, J. L. Hodges, "Discriminatory analysis-nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989, doi: 10.2307/1403797.

[40] I. Khalil, S. Bagchi, "Stealthy attacks in wireless *ad hoc* networks: Detection and countermeasure,", *IEEE Transactions on Mobile Computing*, vol. 10, no. 8, pp. 1096–1112, 2011, doi: 10.1109/TMC.2010.249.

[41] D. I. Urbina, J. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, et al., "Limiting the impact of stealthy attacks on Industrial Control Systems," *Proceedings of the ACM Conference on Computer and Communications Security*, 2016, pp. 1092–1105 [Online], Avaiable https://users.soe.ucsc.edu/~alacarde/papers/ccs16.pdf. [Accessed: Sep. 28, 2022].

[42] B. Filkins, D.Wylie, A. Dely, "Sans 2019 state of ot/ics cybersecurity survey," SANS Institute, 2019.

[43] M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," *Proceedings of the ACM Conference on Computer and Communications Security*, 2016, pp. 1528–1540 [Online]. Available: https://users.ece.cmu.edu/~lbauer/papers/2016/ccs2016-face-recognition.pdf. [Accessed: Sep. 28, 2022].

[44] P. Dash, M. Karimibiuki, K. Pattabiraman, "Out of control: Stealthy Attacks against Robotic Vehicles Protected by Control-based Techniques," *ACM International Conference Proceeding Series*, 2019. pp. 660–672.

[45] J. Li, Y. Yang, J. S. Sun, K. Tomsovic, H. Qi, "Conaml: Constrained adversarial machine learning for cyber-physical systems," *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 2021, pp. 52–66 [Online]. Available: https://par.nsf.gov/servlets/purl/10314482. [Accessed: Sep. 28, 2022].

[46] I. Niazazari, H. Livani, "Attack on Grid Event Cause Analysis: An Adversarial Machine Learning Approach," 2019, doi: 10.48550/arxiv.1911.08011.

[47] E. Anthi, L. Williams, M. Rhode, P. Burnap, A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," *Journal of Information Security and Applications*, vol. 58, no. 8, pp. 102717, 2021, doi: 10.1016/j.jisa.2020.102717.

[48] M. Rajpal, W. Blum, R. Singh, "Not all bytes are equal: Neural byte sieve for fuzzing," *arXiv preprint arXiv:1711.04596*, pp. 1–10, 2017.

[49] J. Li, B. Zhao, C. Zhang, "Fuzzing: a survey," *Cybersecurity*, vol. 1, pp. 1–13, 2018, doi: 10.1186/s42400-018-0002-y.

[50] Y. Wang, P. Jia, L. Liu, C. Huang, Z. Liu, "A systematic review of fuzzing based on machine learning techniques," *PLoS ONE*, vol. 15, no. 8, pp. 1–37, 2020, doi: 10.1371/journal.pone.0237749.

[51] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, et al., "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," *2018 IEEE Symposium on*

Security and Privacy (SP), pp. 19–35 [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8418594. [Accessed: Sep. 28, 2022].

[52] D. Yacchirema, J. S. de Puga, C. Palau, M. Esteve, "Fall detection system for elderly people using IoT and ensemble machine learning algorithm," *Personal and Ubiquitous Computing*, vol. 23, no. 5–6, pp. 801–817, 2019, doi: 10.1007/s00779-018-01196-8.

[53] T. Takahashi, Y. Kadobayashi, "Reference ontology for cybersecurity operational information," *The Computer Journal*, vol. 58, no. 10, pp. 2297–2312, 2015, doi: 10.1093/comjnl/bxu101.

[54] MITRE. *Lateral Movement for Enterprise* [Online]. Available: https://attack.mitre.org/tactics/TA0008/. [Accessed: Sep. 28, 2022].

[55] MITRE. *Lateral Movement for Mobile* [Online]. Available: https://attack.mitre.org/tactics/TA0033/26. [Accessed: Sep. 28, 2022].

[56] MITRE. *Lateral Movement for Industrial Control Systems* [Online]. Available: https://collaborate.mitre.org/attackics/index.php/Lateral_Movement. [Accessed: Sep. 28, 2022].

[57] M. Chorась, M. Pawlicki, R. Kozik, "The feasibility of deep learning use for adversarial model extraction in the cybersecurity domain," *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, 2019, pp. 353–360.

[58] X. Yuan, L. Ding, L. Zhang, X. Li, D. Wu, "Es attack: Model stealing against deep neural networks without data hurdles," *arXiv preprint arXiv:2009.09560*, 2020.

[59] G. Hinton, O. Vinyals, J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[60] S. Kariyappa, A. Prakash, M. K. Qureshi, "Maze: Data-free model stealing attack using zeroth-order gradient estimation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13814–13823.

[61] S. Ghadimi, G. Lan, "Stochastic first- and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013, doi: 10.1137/120880811.

[62] H. Yu, K. Yang, T. Zhang, Y.-Y. Tsai, T.-Y. Ho, et al., "Cloudleak: Large-scale deep learning models stealing through adversarial examples," *NDSS*, 2020. doi: 10.14722/ndss.2020.24178.

[63] L. Zhang, G. Lin, B. Gao, Z. Qin, Y. Tai, J. Zhang, "Neural model stealing attack to smart mobile device on intelligent medical platform," *Wireless Communications and Mobile Computing*, vol. 2020, doi: 10.1155/2020/8859489.

[64] I. H. Witten, E. Frank, M. A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques," *4th Edition, Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann*, Amsterdam, 2017 [Online]. Available: http://www.sciencedirect.com/science/book/9780123748560. [Accessed: Sep. 28, 2022].

[65] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1023/A:1022643204877.

[66] S. Lee, K. Levanti, H. S. Kim, "Network monitoring: Present and future," *Computer Networks*, vol. 65, pp. 84–98, 2014, doi: 10.1016/j.comnet.2014.03.007.

[67] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, et al., "Ensemble adversarial training: Attacks and defenses," *6th International Conference on Learning Representations*, 2018, pp. 1–20 [Online]. Available: https://floriantramer.com/docs/papers/iclr18ensemble.pdf. [Accessed: Sep. 28, 2022].

Utopia Lost – Human Rights in a Digital World

Aaron Brantly | **Department of Political Science, Tech4Humanity Lab, Hume Center for National Security and Technology, Virginia Tech**, USA, ORCID: 0000-0003-4193-3985

Abstract

The long progress towards universal human rights is regressing. This regression is pronounced within digital spaces once thought to be potential bulwarks of a new era in human rights. But on the contrary, new technologies have given rise to threats that undermine the autonomy, empathy, and dignity of human beings. Early visions of human rights being strengthened by networked technologies have instead crashed into technological realities which not only fail to advance human rights discourses, but rather serve to actively undermine fundamental human rights in countries around the world. The future of human rights is increasingly threatened by advances that would make George Orwell blush. Omnipresent data collection and algorithmic advances once promising a utopian world of efficiency and connection are deeply interwoven with challenges to anonymity, privacy, and security. This paper examines the impact of technological advances on the regression of human rights in digital spaces. The paper examines the development of human rights through changes in concepts of autonomy, empathy, and dignity, it charts their regression as technologies are used to increasingly prey on these very same characteristics that undergird human rights discourses.

Keywords

artificial intelligence, cybersecurity, governance, human rights, privacy

Corresponding author:
Aaron Brantly, Department of Political Science, Tech4Humanity Lab, Hume Center for National Security and Technology, Virginia Tech, USA, ORCID: 0000-0003-4193-3985; abrantly@vt.edu

Cite this article as: A. Brantly, "Utopia Lost – Human Rights in a Digital World," ACIG, vol. 1, no. 1, pp. 34–55, 2022, DOI: 10.5604/01.3001.0016.1238

1. Introduction

Human rights as a concept have progressed substantially over the last 100 years. From the Universal Declaration of Human Rights (1948), the signing of the post war Geneva Conventions (1949), and multiple national constitutions, and laws enacted by states and unions of states there has been a drive to expand and protect human rights. The recognition and protection of these rights has proceeded unevenly over the last 80 years. The inconsistent safeguarding of human rights has been undermined or ignored in a variety of political and social contexts in nearly every state. Yet where these rights seek to establish a robust ground upon which to base fundamental rights inherent to all humans, the advances of networked technologies have provided a new and pervasive means by which states can, with high degrees of efficiency, erode rights once normatively and legally established. Networked technologies that were once believed critical to ushering in a more just world respectful of the rights of human beings, have instead been co-opted to surveil, censor, and constrain rights. While utopia was never a reality, this loss of rights, and the shifting of the normative frameworks on human rights, represents a lost vision of a utopia in which the rights and dignity of humans could have been. The analysis below is constrained to digital human rights violations whose attributes undermine human autonomy, empathy, and dignity.

Alexander Wendt is famous for his constructivist turn of phrase "anarchy is what states make of it"[1]. Similarly, the development of the norms and ideas surrounding the conceptualization of human rights are what states make of them. Nearly a century of progress towards the rights of human beings is being undermined through the slow alteration of the ideational and normative constructs about what constitutes rights and who should and can respect those rights. Whereas the development of human rights followed an often painfully slow process of norm evolution through an ever-progressing norm life cycle [2], that cycle never completed, and the internalization of human rights norms in nearly all states has begun to reverse itself. This reversal was forecast by a few scholars, notably by Ron Deibert in a series of volumes examining the encroachment of the state through the utilization of digital means to undermine human rights [3–5]. Subsequent research on the rate of change indicates that the speed of reversion away from a recognition of human rights in digital spaces correlates with the rate of change in digital capabilities developed by states [6]. Early concerns about the impacts of technological advances in networked technologies centered on authoritarian regimes [7]. Despite informed warnings about authoritarian counter movements utilizing technology to undermine the advances of norms on human rights, many in the academic and policy communities professed a profound and not entirely unwarranted optimism about technology and its power to enable human rights[1]. Among the scholars who led both the academic and policy discussions on the ability of technology to facilitate human rights was Larry Diamond, who in writing on the application of technology to civil and political rights spoke of the potential for technology to "liberate" and empower individuals [8]. The empowerment vision often correlated with technological advance is not without merit. There is substantial evidence that networked technologies

1 —— A large volume of research is available on the early positive benefits and some challenges associated with new technologies and human rights. The works listed here are but some of many that express substantial optimism about the role technology can play in facilitating human rights [110–115].

enabled social mobilizations to challenge authoritarian and rights abusing states and state institutions [9, 10]. Yet these challenges were often met by the counter usage of networked technologies for highly repressive and intrusive digital surveillance and manipulation [11].

Ron Deibert and the Citizen Lab at the University of Toronto have been instrumental in identifying and bringing to public consciousness a variety of violations of human rights norms [12, 13]. Deibert in particular, has been outspoken in highlighting what he identifies as the need for a "reset" in the balance between, on the one hand, implementation and use of technology, and on the other, human needs [14]. The stories of human rights and digital rights have not transpired in isolation. They are intrinsically enmeshed. Organizations ranging from Amnesty International, Human Rights Watch, Doctors without Borders, and others have increasingly joined digitally oriented rights organizations such as the Electronic Frontier Foundation, Access Now and many more in a common push to secure human rights in digital spaces. In truth, rights defended in digital spaces are not meaningfully distinct from those same rights expressed in non-digital spaces. And very often violations of rights in digital spaces occur in tandem with human rights violations in physical spaces.

Despite human rights violations in digital and physical spaces being highly correlated, the rights within the two spheres do not carry with them the same normative value. The result is that rights, once freely exercised through digital means, are increasingly undermined as state capacity to control digital spaces has increased. Yet the restriction of rights is not solely the result of states recapturing rights once previously held. They are aided in their capture by a range of actors who, using the market and the means of surveillance capitalism [15], alter the norms associated with digital human rights more broadly. The result is a subtle yet profound shift in concepts of free speech, privacy, and surveillance to name just a few of the broader spectrum of rights impacted.

This work contextualizes the formation of digital human rights within the larger and comparatively more robust history of human rights outside of digital spaces. Examining the construction of norms associated with the formation and subsequent decline of rights in digital spaces through a constructivist lens, this work answers the questions 'How?' and 'Why?' digital rights are regressing despite increasing advances in networked technologies once heralded as tools of human rights empowerment. The work proceeds below in four sections. The first defines both human rights and norms, it provides a brief history of the construction of human rights norms. The second section examines the rise of digital rights and the utopian views associated with rapidly advancing networked technologies. The third section examines the decline of rights in digital spaces in the context of a failure to solidify norms around digital rights in relation to human rights. The final section provides a discussion on the loss of rights and the path forward for digital rights norm entrepreneurs.

2. Constructing Conventional Human Rights

Human rights are a modern concept within the Western political cannon. On the origins of human rights Lynn Hunt notes "Human rights are difficult to pin down because their definition, indeed their very existence, depends on emotions as much as on reason" [16]. Hunt's implication of the emotional attributes resident in human rights is indeed central to what amounts to a constructivist argument which she develops over the course of her work. She isolates a core tenet of human rights that this work seeks to develop in greater detail, concepts of perception concerning both the self and other, and the recognition of a simultaneous uniqueness and universality of thought and condition. The development of these concepts into an applicable and meaningful body of cultural and societal knowledge and identity is constitutive. Hunt specifically and parsimoniously identifies the concepts of autonomy and empathy as critical to the constitution of norms on what would eventually evolve into human rights. It is important to note that human rights framed in such a way as to privilege the autonomy of individuals, i.e. individualism, is not culturally universal and has implications in non-western societies. Yet, to set a starting point, this paper emphasizes the concepts of individuality and empathy as a basis for understanding how human rights are conceived in physical spaces, and how these same rights fail to carry over into digital spaces. It is also important to underscore advances in philosophical understandings of human life and value. In particular, and often related to the notion of individual autonomy and empathy, is the concept of dignity outlined by Immanuel Kant [17]. Kant bridged the concept of autonomy with dignity in writing "Autonomy is therefore the ground of the dignity of human nature and of every rational nature." [18]. Yet a deepening understanding of the concept of dignity into the broader field of human rights did not occur quickly.

Converting autonomy, empathy, and dignity from abstract concepts into codified legal structures was not straightforward. While enlightenment thinkers debated concepts of humanity, moral and ethical behaviors, and western authors and artists probed the mind of the individual and their unique visual appearance [16], these concepts were in opposition to millennia of lived and learned experience. Constructing identities encompassing such concepts required shifts in social, cultural, economic, political, religious, and other framings. Two areas that helped to facilitate new identities were shifts in both philosophical and artistic works which served as something akin to fuzzy norm entrepreneurship. I deliberately use the term fuzzy[2] because unlike many modern norm entrepreneurs the concepts surrounding human rights were not codified in a manner that allowed for specificity, nor were they advanced in most instances by a single group. Rather there remained only the notion that the order as it existed was not as it might be.

Specific political philosophers, such as Mary Wollstonecraft [19], pushed back on the early formulations of rights assigned to men implicitly or explicitly in law, and implicitly in works designed to further the formation of new identities rooted in rights-based discourses [20, 21]. Despite a fervent discourse that permeated reading circles on both sides of the Atlantic, initial implementations of these identity constructs for rights were entirely focused on emphasizing the propertied white male class. The most famous

2 —— My use of fuzzy norms contrasts with that of [116] in which states deliberately fail to define the parameters of a norm. By contrast norms here are "fuzzy" simply because they have not been articulated in a universally applicable manner in line with current cultural or societal power structures.

documents in the modern western cannon, such as the US Declaration of Independence, formally established "rights" in official documentation through the words: "We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness"[3]. These words expressed the ideals of changing social and cultural norms within existing entrenched power hierarchies. The result was the generation of rights for some and the often-brutal withholding of rights from others[4]. The concept of privileged access to rights will return in later sections. Subsequent attempts, such as the Declaration of the Rights of Man and of the Citizen (1789), similarly advanced broader concepts pertaining to human rights, yet their implementation, both in domestic environments and in colonial possessions, remained highly entrenched in systems that deliberately sought to deny rights to overwhelming majorities of inhabitants.

The fundamental development of human rights as a concept has historically also privileged the political and social rights of individuals rather than basic economic rights[5]. The separation of the economic from the social and political plays a part in the story of the regression of rights in the digital age. Yet it also influences a broader understanding of human rights in the western political context. The development of human rights is nonlinear, just as all normative advances are by and large non-linear. The literature is replete with examples of the asymmetric application of human rights based on any number of factors encompassing race, class, religion, and many other attributes. Eric Weitz's *A World Divided highlights many of the inherent tensions and juxtapositions of rights within a range of communities from Namibia to Minnesota, to Brazil and Haiti* [22]. Weitz identifies that when individuals are not conceived of as belonging to a state, with citizenship, they have been historically exposed to the worst forms of degradations [22]. Arendt, writing at the end of the Second World War, identified being "stateless" as one of the worst conditions imaginable, a condition only marginally better than physical annihilation [23].

It took the cataclysm of the Second World War to begin altering social and political reality towards a more universalized understanding of Human Rights. The Universal Declaration of Human Rights (UDHR) ratified on December 10th, 1948, was pushed forward by norm entrepreneurs led by Eleanor Roosevelt in the aftermath of millions of deaths at a time when many European powers were advancing towards or experiencing colonial collapse. In many ways the UDHR arrived at what John Kingdon would refer to as a "policy window" considered as an important opening for agenda setting [24]. Work undertaken by activists, political philosophers, politicians, and many other actors capitalized on the human tragedy of war to draw together a codified understanding of human rights. Lest there be any doubt as to the reality of such a policy window being fostered by the tragedy of war, it is important to remember that prior to World War II the term "human rights" was not used with any measurable frequency [25].

The first article of the UDHR plainly states: "All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood" [26]. This statement, while inherently gendered in its formulation, ties together the concepts of autonomy, dignity, and empathy. The application of

3 —— https://www.archives.gov/founding-docs/declaration-transcript.

4 —— In the case of the US Articles of Confederation and later the US Constitution numerous caveats are made to exclude persons and reduce both their individuality – independence and uniqueness – black men were considered 3/5s a man, women and non-propertied males were excluded.

5 —— See for a detailed analysis of the tradeoffs associated with an emphasis on political and social rights rather than economic rights [117].

these rights, while declared universal, were in practice selective[6]. Moreover, the protection of the outlined rights was left to states. The concept of dignity as expressed in the UDHR is best explained by political philosopher Jürgen Habermas, who writes that human dignity serves as the "conceptual hinge" linking the "internalized rationally justified morality anchored in the individual conscience" of Kant and the "coercive, positive, enacted law" of modern states [27]. The UDHR constitutes the output of hundreds of years of progress in fusing recognition of the autonomy, empathy, and dignity of human beings with the legal mechanisms of states capable of enforcement. Despite its monumental success in establishing a relative normative consensus on what constitutes human rights, the UDHR was not enforceable without the consent of nations. Where normative consensus arose conceptually, the translation of the concepts to applications was undertaken in piecemeal fashion by nations through their "obligations" to the ideals established in the UDHR.

This state centric approach to human rights is rooted in the sovereign rights of states established in a post-Westphalian order and presents a paradox for the protection of rights. The principal violator of rights is also obliged to be the protector of those same rights [25]. Unlike the fuzzy formation of concepts pertaining to autonomy, empathy, and dignity, developed over nearly two centuries, the emergence of increasingly powerful norm entrepreneurs capable of forming both state and non-state-based organizations to further advance norms pertaining to human rights occurred in the period after the signing of the UDHR. Although the broadest and furthest reaching statement of human rights was the UDHR, it was quickly followed by the Genocide Convention (1948), Refugee Convention (1951), Discrimination in Employment Convention (1960), Racial Discrimination Convention (1965), the Economic, Social and Cultural Rights Covenant (1966), the Civil and Political Rights Covenant (1966), the Discrimination against Women Convention (1979), the Convention against Torture (1984), the Children's Convention (1989), the Indigenous and Tribal Peoples Convention (1989), the Convention on Migrant Workers (1990), the Convention on Persons with Disabilities (2006), and the Declaration on the Rights of Indigenous Peoples (2007). This multitude of conventions and declarations arose in tandem with a rapid increase in organizations dedicated to fighting for and exposing human rights violations globally as well as within nations. Although the first major international rights organizations such as the International Federation of Human Rights (FIDH) predate the UDHR, it was in the post-World War II era where organizations such as the FDIH, the International Commission of Jurists (1952), and later Amnesty International (1961), Human Rights First (1978), Human Rights Watch (1978), and others began to emerge and exert increasing pressure on states to adhere to their commitments under international law.

Although progress has been made on both conceptualizing and enforcing human rights norms, norm violations are a regular occurrence. A substantial body of literature has examined the development of norms pertaining to human rights [28–31]. The findings are mixed, they provide a range of explanations as to how early norms on human rights emerged [28, 32], why states violate human rights [31], and why they adhere to human rights [31]. Yet despite many failings, there has been substantial progress

6 —— Many states sought exemptions or explicitly denied rights to persons within their jurisdictions. This was true in practice with nearly every western state which became a signatory to the UDHR.

on human rights norms [33]. The evolution from Hunt's historical framing of autonomy and empathy, to Kant's initial definitions of human dignity, onward to the flawed and often hypocritical declarations of "Independence" and the "Rights of Man" to the emergent post war policy window making possible the codification of a substantial foundation for human rights in the UDHR, has been followed by 70 years of rapid norm emergence up to the present. This situation has been fostered by an ever-growing cohort of states and NGOs that have raised established human rights as being integral to the lexicon of international politics. Imperfections and failings abound. Yet the literature, and the wider policy discourse on human rights, indicate that it has found a steady body of norm entrepreneurs willing to continue advancing it forward often in the face of great hardships.

In the last 30 years a subset of human rights emerged as an addendum to those rights previously fought for and often enumerated. Digital rights once discrete and thought the purview of a select few in connected western nations have increasingly become intertwined in all forms of human rights. Digital Rights are human rights and in the next section I briefly trace the emergence of norms and challenges emanating through networked technologies. These norms are framed in such a way as to identify how they impact the concepts of human autonomy, dignity, and empathy. Constructing the framing of norms around these concepts is rooted in an ontological notion, i.e. the state of being of rights, outlined above, leading to the establishment of rights that form a recognition of individual human autonomy, both internally and externally identified and respected through empathy, and codified in the complex concept of dignity developed by both Kant and Habermas. Autonomy and empathy are constituent parts of the larger meta-concept of human dignity. When combined these three concepts underpin the creation of what are considered human rights. Understanding how technologies influence both the constituent components and the broader overarching concept establishes how human rights are affected by digital technologies.

3. The Rise of Digital Rights

The history of the Internet and its associated technologies has been well researched [34, 35]. The transition from circuit-switched to packet-switched data [36], followed by advances in networking large and expensive computers with Interface Messaging Processors (IMPs) [37], and eventually protocol and software suites such as the Transmission Control Protocol/Internet Protocol (TCP/IP) [38] is a socio-technical story of development that occurs within both civilian and military environments. The development of networked technologies is imbued with the hopes and fears, constraints and freedoms associated with the times in which it was initially developed. Early governance of the Internet was partially conducted through the development of the Request for Comment (RFC) process which reinforced a technocratic approach that was later to be enshrined in the nascent but developing governance structures of the Internet, including what has now become the Internet Society. The technocratic nature of the Internet often overlooked or under-estimated its expanding power and reach.

Early Internet development was hamstrung by government regulation, both from the National Science Foundation [39] and the International Traffic in Arms Regulations (ITAR) managed by the Department of State [40]. ITAR restrictions pertaining to the use of the evolving network were particularly contentious, they dealt with what can be best described as an early debate over individual rights in digital spaces. The debate led to what is commonly referred to as the first crypto war [41]. During the crypto wars of the early 1990s members of the Intelligence community and in particular the Federal Bureau of Investigation (FBI) took a strong position against allowing the commercial use of cryptography [42]. At the time of the initial fight the justification against allowing public use of cryptography centered on the role of the state in accessing private information. It is important to note that at the same time the US was debating the de-listing of cryptography from ITAR, the Communications Assistance for Law Enforcement Act was being pushed forward in the US congress, meanwhile the National Security Agency in coordination with other law-enforcement agencies were pushing the introduction of the "Clipper Chip" to provide a secure backdoor into all US digital communications [43]. Additional constraints on the developing network and its legitimacy arose from its addressing architecture which centralized control into the hands of one person, Jon Postel. Such control was later exercised by the International Corporation for Assigned Named and Numbers (ICANN) under the National Telecommunications and Information Administration within the Department of Commerce [44].

As expanding Internet infrastructure increasingly made possible robust decentralized communication, the fight over who would control this communication was just getting started. Early utopian norm entrepreneurs such as Grateful Dead lyricist, John Perry Barlow, saw the fight as intimately related to rights [45]. As the conventional human rights community was regaining its footing after the 1980s and dealing with large geopolitical changes related to the collapse of the Soviet Union, concerns pertaining to the protection of rights in online spaces largely fell to the wayside. Yet some NGOs did arise and fought to include digital considerations in broader discussions of human rights. Two prominent US based organizations, the Electronic Privacy Information Center (EPIC) (1994) and the Electronic Frontier Foundation (EFF) (1990), were established to defend civil liberties in the digital world. Concurrent to these formal norm entrepreneurs pushing forward or combating various forms of legislation, informal norm entrepreneurs in many countries built increasingly robust groups of hackers that challenged the status quo of what it meant to have rights in digital spaces. Groups including the Chaos Computer Club (1981) [46], Cult of the Dead Cow (1984) [47], L0ft (1992), and numerous others, expanded global interest into digital rights through cultural events, hacktivism and collective organization. Early language on digital rights sought to draw direct relationships between rights in online spaces and physical spaces.

Whereas the formation of rights leading up to and including the UDHR took several hundred years, early digital rights activists tried, and in many cases, succeeded in tying rights in one space to rights in the other. The passionate community of hackers combined with legal and policy wonks to foster robust dialogues on topics ranging from the vulnerabilities in

government backdoors [48], to fundamental concerns pertaining to privacy [49]. These efforts have generated a range of governmental and non-governmental responses. Efforts such as the Internet Governance Forum and Multiple United Nations Governmental Groups of Experts, attempts by the International Telecommunications Union and others, have sought to raise to international attention critical issues pertaining to digital spaces. These efforts have spawned contentious debates on the role of state and non-state actors in the governance of the Internet [50, 51]. They have raised issues of multi-lateral versus multi-stakeholder involvement in how the Internet functions and what rights and privileges of individuals are to be protected and by whom. Internet governance debates do not lack norm entrepreneurs seeking to shape the identity of networks and netizens.

Although organizations including EPIC, EFF, Access Now, European Digital Rights, Digital Rights Watch, Internet Freedom Foundation, Fight for the Future, and several others, have increasingly sought to tie global Internet governance concerns to individual issues, these issues are often drowned out in technocratic and bureaucratic discussions. In particular, the technocratic push towards a future "utopia" of digital artefacts often undermines basic human rights normatively established and largely agreed to outside of digital spaces. Technocratic and market incentives are driving a divergence away from normative advances in human rights outside of digital spaces and resulting in a regression of digital human rights.

The next section examines the regression of human rights in digital space through both technical and policy lens and examines why norms of rights so robustly established outside of digital spaces are under so much threat within them.

4. Losing Utopia: The Regression of Digital Human Rights

A once quasi-anarchic, libertarian leaning space filled with hackers and NGOs fighting for "independence," free flows of information [52] linking societies around the globe in a generative [53] and collective march towards a better vision of efficiency and connectivity [54] is now increasingly contested. The norm entrepreneurs fighting for digital rights as human rights have not gone away. If anything, the number and scale of norm entrepreneurs fighting for digital rights has increased globally. Organizations traditionally focusing on human rights, democracy, and the rule of law in physical spaces, such as the National Democratic Institute, the International Republic Institute, and many others increasingly added to their portfolio digital rights. In the early 2000s under then Secretary of State, Hillary Clinton, digital rights defenders even found common cause with parts of the US government [55]. This common cause did not last long. In mid-2013 former NSA contractor Edward Snowden began releasing large volumes of documents through a variety of media outlets demonstrating the reach and extent to which the US government was capable of undermining digital rights online [56, 57].

How was utopia lost? First, it is important to note that the vision of utopia expressed by many academics, policymakers, and corporations did not exist in the way it was often portrayed. Notable pushback arose during

this same period with some scholars applying a severely critical lens to the utopian visions being professed [58]. Moreover, there was substantial early analysis that suggested unease within much of the non-democratic world over the influx of new connected communications technologies [7]. Rebecca MacKinnon's efforts to document the multitude of violations arising from the influx of digital technologies illustrates that the present often regressive state of digital freedoms arose from a continually contested understanding of rights in digital spaces. Substantial evidence presented by Philip Howard and Muzammil Hussain underpinned the reality that states have been engaged in substantially repressive behaviour since networks started expanding outward in the late 1990s and into the early 2000s [59, 60]. In particular, Howard and Hussain write of the process through which new technologies were introduced into states and then subsequently repressed:

> A preparation phase, involving activists' use of digital media across time to build solidarity networks and identification of collective identities and goals; an ignition phase, involving symbolically powerful moments which ruling elites and regimes intentionally or lazily ignored, but which galvanized the public; a protest phase, where, by employing offline networks and digital technologies, small groups strategically organized on large numbers; an international buy-in phase, where digital media networks extended the range of local coverage to international broadcast networks; a climax phase, where the regime maneuvered strategically or carelessly to appease public discontent through welfare packages or harsh repressive actions; and finally, a follow-on information warfare phase, where various actors, state-based and from international civic advocacy networks, compete to shape the future of civil society and information infrastructure that made it possible [59].

The preparation, ignition, protest, international buy-in, climax, follow-on information warfare chain is in many ways an expanded understanding of norm dynamics presented by Finnemore and Sikkink in their work on international norm dynamics [2]. Early norm entrepreneurship for digital rights emphasized the spread of information communications technologies to countries along with the value such networks would bring with them [61]. Early internet freedom advocates saw the value of networks as tools for advancing human rights, many also recognized the subsequent repressive activities of states in response. The value of the networks in advancing technologies spawned multiple use case specific technologies to facilitate democracy and human rights. Projects such as the Guardian Project, the Tor Project, Tails, Cryptocat, and many others, provided a means for democracy and human rights activists to protect their data from intrusive states. These applications built on rapid developments in the open-source software and cryptography communities to enable network development in the later stages of the cycle proposed by Howard and Hussain when states increasingly used repression. Yet the reality remains, as demonstrated by the Snowden releases and reports of increasingly powerful malware developed by state and private corporate firms, that the ability to ensure human rights in networked spaces were increasingly under sustained threat [62–64].

Norm entrepreneurs for human rights in digital spaces have faced an ever-increasing array of challenges since 2010. Where once state-based actors were the principal threat to digital rights, the threat landscape has become increasingly complex as the technological landscape has shifted towards big data, machine learning, artificial intelligence [65], the Internet of Things (IoT) [66], and social networks [67–69]. At the forefront of the shifting vocabulary of norms on rights in digital spaces have been large technology companies [15]. The shift in vocabulary has positioned violations of human rights as consequences of technological advances or temporary setbacks resulting from flaws in code or algorithmic design. Yet the systematic and pervasive penetration of technology into every facet of daily life in nearly every country around the world comes with profound consequences for the development of human rights in digital spaces. Disaggregating the attributes of the regression of human rights norms in digital spaces is complex as many technologies that have facilitated regression overlap and foster human rights challenges in divergent forms in different societies. In much the same way as conventional human rights violations occur and are addressed differently in different states, the same holds true for digital human rights violations.

As stated in the introduction, this paper cannot address all possible digital human rights violations, so instead it seeks to address those attributes that undermine human autonomy, empathy and dignity. Changing norms in digital spaces erode concepts of human autonomy, empathy, and dignity and therefore strike at the heart of normative discourses on digital human rights, they also serve as the fuel for digital rights regression. The regression of concepts pertaining to autonomy, empathy, and dignity arise from advances in data collection and data usage. Data collection can be further subdivided into either direct or passive interaction, while data usage can be subdivided into algorithmic (ML/AI) applications, hybrid, and individually targeted applications.

Data collection is no longer relegated to interactions of individuals in perceived digital space. Most users perceive digital interactions as originating from direct engagement with digital artifacts such as web-browsers, search engines, or other forms of active engagement. While early interactions in digital spaces were principally the result of direct interaction, the move from Web 1.0 to Web 2.0 and beyond has increasingly shifted the vast quantity of data interactions from active to passive engagements [70–72]. Examples of passive data collection abound. Presently individuals carrying mobile phones are constantly providing geolocation data with or without GPS settings activated. As individuals move between cell towers their relative position is relayed to the mobile provider. This relative position is tied to a device that in many countries is also tied directly to an individual's identity. The physical movement of the device in proximity to a person can provide data on whether an individual was in the vicinity of a perpetrated crime [73] or can be used to identify individuals engaged in protest [11].

Mobile devices are frequently used as navigational aids to assist drivers or other travellers as they move between destinations. These devices in turn provide substantial data on everything from user interests when stopping at stores or gas stations, to speed and telemetry data

used at the individual level to monitor driving behaviour, or in aggregate to assess traffic patterns [74]. Passive collection data extends from the devices carried to individual level telemetry data in stores and public spaces through to use of Bluetooth protocols [75]. Individual level tracking mechanisms such as these are meant to provide greater efficiency to users and customers. They improve business efficiency and help facilitate the sale of advertisements. Yet in the process of providing data through constant passive interactions the individual loses autonomy. What at first glance appears to be pure gains in efficiency in turn is the conditioning of individuals through repeated passive interactions to shape and orient behaviour. Traffic guidance applications present notifications portraying advertisements to nearby shops, reroute traffic to avoid congestion through neighbourhood streets, and provide insurance companies with data to adjust automotive insurance rates [76]. While the above examples principally originate in the private sector, the technologies used are universal and governments around the world have increasingly relied on telemetry data derived from passive collection to facilitate state-based repression [77].

Violations of privacy are directly related to passive data collection and autonomy of self. Autonomy Privacy is defined as "an individual's ability to conduct activities without concern of or actual observation"7. Twenty to thirty years ago passively divulging information of a sensitive and personal nature to individuals would likely have been considered a substantial violation of privacy. Yet with the advent of tools capable of enabling passive data collection on individuals in nearly every aspect of their lives autonomy privacy has been eroded [78]. The intrusion of the digital world by both corporate and state actors is in direct violation of Article 12 of the UDHR:

> No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks[8].

8 —— https://www.un.org/en/about-us/universal-declaration-of-human-rights

Yet in the span of 20 years what was once a human right has become a privilege, this is so because to secure the human right to privacy one must exclude oneself from modern social and economic infrastructures. Even a full exclusion of oneself from platforms and devices is not a guarantee of privacy. Social Media firms such as Facebook have found ways to collect data on individuals who do not even subscribe to their platforms [79]. In China new efforts to establish national facial recognition has resulted in the mass passive collection of data for state and corporate uses. The subsequent utilization of these data is virtually limitless [80]. Following in China's footsteps, India has begun the process of collecting and developing a national biometric database on its entire population [81]. Passive and active data collection are only likely to increase, this will result in further gross violations of privacy.

Data collection in online platforms is used both within and beyond digital spaces to undermine an array of human rights. Notable examples arise in China, which leverages big data to assess when online movements veer towards collective actions which might challenge the state or undermine state narratives [82]. The use of data collected from both

passive and active engagements undermines the original utopian visions of the Internet as a platform for overcoming collective action problems [83]. Passive data collection even extends into private educational spaces as schools and universities have increasingly leveraged passive data collection to create models that can be used in later AI and ML applications to accuse students of cheating on digitally administered tests [84]. As private and state capacity to control data increases the autonomy of individuals and groups declines. This undermines several articles articulated under the UDHR, most notably articles 19 and 20[9]. Some norm entrepreneurs have had success in pushing back against the "more data is better" argument of business and many governments.

The General Data Protection Regulation which came into effect in 2018 is the principal example of a success which contrasts trends seeking to further erode rights in online spaces [85]. Implementation of the GDPR has had cascading impacts outside of European Union jurisdiction. Most easily recognizable of the impacts is the notification requiring users to accept the collection of cookies when visiting websites that might have users from the European Union. But just as this is an example of success, it is also an example of the embedded problems associated with safeguarding digital rights. The notification to accept cookies, instead of being a protective measure, is instead a further mechanism normalizing the collection of data through "user consent" [86, 87]. Just as most users fail to read the terms and conditions for most products and services due to their length and linguistically obtuse language, so too do users accept tracking cookies on web platforms. Although the GDPR has provisions safeguarding the rights of individuals in a digital environment the implementation of these rules is complex and requires constant oversight. Violations are punished principally through monetary fines.

Moving beyond discussion of data collection, the algorithmic manipulation of data erodes autonomy, undermines empathy, and imperils dignity. Just as most users are unaware that data is constantly being collected through passive interactions, they are also largely unaware of the impact that algorithms have on individual decisions. The manipulation of human decisions using algorithms occurs in the background, it is obscured via user interfaces and claims that these algorithms save time or simplify decision making processes. One of the most blatant examples of a private organization using algorithms for behavioural manipulation arose when Facebook undertook an experiment in which they presented only negative content in more than 700,000 user feeds [88]. The intent of the experiment was to generate an "emotional contagion" and by all accounts their experiment was a success. As a result of altering the algorithm Facebook was able to shift the emotional disposition of users. This power has profound human rights implications that are not well articulated within the existing UDHR. Facebook acknowledged in internal reports that the ability to successfully manipulate users could potentially alter democratic elections [89]. Such algorithmic manipulations demonstrably alter the notion of individual autonomy. Worse still, the use of algorithmic manipulation falls outside of the typical target of human rights obligation – governments – and instead empowers private actors to violate human autonomy as a result of market factors.

9 —— Article 19 states: "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers." Article 20 states: "
1. Everyone has the right to freedom of peaceful assembly and association.
2. No one may be compelled to belong to an association."

Issues of autonomy highlighted by the Facebook emotional contagion study are not the only area of concern. Manipulation of the psychological state of users, in particular levels of empathy, have been demonstrated to have mixed effects dependent on the structure and formulation of the information environments [90]. Network structures, as conditioned by algorithmic design, can in some instances impact the susceptibility of individuals to certain types of information that might either increase or decrease empathy towards issues and persons outside of one's own experience. Networks with algorithms fostering filter bubbles can alter the disposition of individuals in both positive (rights affirming) and negative (rights denying) manners [91]. The presentation of algorithms as neutral in curating the digital artefacts or in providing information to users is far from the truth. While software (of which algorithms of ML and AI are a subcategory) has often been perceived as agnostic politically, culturally, racially, and economically, it is in fact an expression of power intentionally or unintentionally constructed. Computer code, the instructions on which digital systems run, is the base construction of algorithms that make social media platforms, applications, and search engines function, it is a technical design with imbedded social and cultural values [92]. These values are not neutral. As a result the implementation of ML and AI not only reduces human autonomy but alters the empathy of users in ways that change their perception of rights and their perception of others.

Where Hunt illustrates how literature and art foster common humanity, Safiya Noble demonstrates that algorithms can have the opposite effect and result in the dehumanization of individuals [93]. Her work unequivocally demonstrates how platforms such as Google's search engine undermine the empathy critical to fostering and forming human rights claims within populations. Her work illustrates how platforms ostensibly developed to connect and share information can concurrently marginalize and undermine the political, social, and cultural positions of minorities and underprivileged groups. Expanding beyond bias embedded in the representation of individuals, algorithms have violated the equal right work considerations contained in Article 23 of the UDHR. Amazon and other companies used algorithms in their hiring systems to build profiles on potential employees. The result of these practices was substantial bias and reinforcement of existing labour pool ethnic composition [94]. The role of algorithms in undermining human rights even extends to the borders of nations and includes infringements on everything from privacy to freedom of movement [95].

Algorithms are increasingly pervading every aspect of modern digitally connected life. The implications for human rights violations arising from the development and use of algorithms is substantial and demonstrated [96]. Despite repeated documented failures and examples of algorithms resulting in human rights violations they, like big data collection, largely fall outside the purview of conventional human rights discourses. This is slowly changing as scholars address the legal and regulatory consequences of their implementation in everyday life [97]. Increasingly discussions on the security of individual rights are being combined with concepts of cybersecurity and related topics and themes [98]. However, despite early efforts to reign in algorithms there remains a robust push to advance AI and ML applications without regard to their impact on

digital human rights. Often this push is a function of market mechanisms embedded within surveillance capitalism, but just as frequently these use cases arise from academic scholarship at the nexus of computer science and multiple other fields of inquiry utilizing data and algorithms to solve specific problems. The challenge presented is almost the reverse of that faced by early norm entrepreneurs seeking to foster new conceptualizations of rights. With advances in data and algorithms it is the primacy of technological advance for some hypothesized utopian vision of an efficient and profitable world that overwhelms the rights-based discourse and fosters normative regression.

At the intersection of data and algorithms resides the central challenge to digital human rights. It is at this intersection where the dignity of human beings is undermined, where they are converted from UDHR Article 1 – "human beings born free and equal in dignity and rights" to potential manufacturers of data and consumers of products to be manipulated and directed in systems of digital control. It is here where the humanity of the human being is transferred into bits and bytes to be analyzed, organized, and directed. Data and algorithms are combining to enable technologies that undermine human dignity. Firms such as Cambridge Analytica prey on citizens in multiple countries around the world[10] through their data, and leverage algorithms to enable tailored manipulations to alter the outcomes of elections [69]. Firms manipulate the perceptions of human worth and value through the weaponization of information infrastructure for profit or political gain [99–101]. The constant and increasing challenges associated with data and algorithms impacting on human dignity are likely to grow as they influence everything from employment and education [102], to healthcare [103, 104] and criminal justice [105].

10 — See for example the discussion on the use of data by Cambridge Analytica to undermine Kenyan elections [118].

Big data collection and algorithms are fostering a steady regression in discourses on human rights in digital spaces. Rather than any one technology being presented as a fundamental violation of human rights, the issue at hand is a change in the normative discourse associated with rights secured in and through digital environments. Just as the march towards a common human rights discourse was slow and contested, the movement away from human rights in digital spaces is occurring in a slow, steady progress of technological advances, each attacking a slightly different area of concern. The failure to solidify the discourse around rights in digital spaces means that such rights have been exposed to the pressures of the market and the power of technological advance.

5. Re-establishing Digital Rights Norms

Kieron O'Hara and Wendy Hall present a compelling case that the regression of rights does not occur uniformly in all digital environments [106]. They argue that the political structures within nations in which networks grow and develop heavily influence how those networks are run and the rights associated with those networks [106]. They note that there are five "Internets" emerging globally. Among these are the Silicon Valley model of openness, the Brussels Bourgeois Internet, the DC Commercial Internet, The Beijing Paternal Internet, and the Moscow Spoiler Internet. Each of the

five visions of the Internet in their analysis presents a unique set of policy and regulatory challenges. The argument that the Internet is fragmenting into zones of control and regulation is not new and has received some pushback from scholars such as Milton Mueller [107]. Different states are exerting different levels of control over their domestic Internets. Some states are increasingly repressive, while others are balancing the ills of openness with those of control. Yet what these arguments overlook is the advancing march of technology, in particular its ability to collect and use data in novel ways that strike at the heart of human rights discourses. The future of the Internet is uncertain. It is filled with enormous promise and peril. It is facing a future of information liberation [108] and censorship [109].

The evolution of human rights norms from shifts in autonomy, empathy, and the eventual definition of human dignity were not straightforward. While some individuals, landed white Christian males, experienced improvements in human rights earlier in the norm life cycle, the eventual development of the UDHR and a bevy of other conventions and laws at the national level shifted discourses and have made an impact. While early discourses on rights in digital spaces were strong and heavily supported by certain states, the consistent push for digital rights to be recognized and considered as human rights has suffered as technological advances have increasingly manipulated the conversation away from rights towards conversations on economics, efficiency, scientific advance and more. These counter norms and discourses damage the norms meant to foster autonomy, empathy, and dignity. They obscure their motives and impact with code and hardware. They shift the once quasi-utopian vision of a liberating Internet towards one that constrains rights and freedoms. Whereas the movement towards expanded human rights fought to elucidate and clarify those attributes of humanity that needed protection and from whom humans needed protection, the regression of rights in digital spaces is subtle and opaque. The march of technologies without thoughts to human security and rights approximates to the placing of a frog in a pot of water and slowly raising the temperature until it is boiled.

Digital technologies increasingly impact human autonomy, dignity, and empathy. They alter the way citizens, governments, and firms see and interact within one other. Technologies create dependencies and efficiencies that can and often do harm human security through reductions of human autonomy and the alteration of human empathy. Digital technologies create new means of violating human rights which are exclusively digital, they also extend older violation typologies from physical spaces to virtual ones. Yet of equal importance digital technologies can and do extend from virtual spaces back into physical spaces in ways that profoundly undermine human rights. In many ways digital rights violations are extremely pernicious because they extend into the personal spaces of individuals which were once free from surveillance mechanisms accessible to governments, firms, or even fellow. Rights once explicitly protected, are increasingly subject to terms of service, algorithmic design, pre-digital understandings of previously secured rights, and more. The result is that human autonomy in digital spaces is increasingly not a right, but a privilege secured through either payment to firms, or complex security practices learned and implemented by individuals. Through algorithms, networks, data collection and analysis, and platforms

that shift the perceptions of others technologies are increasingly attacking the foundations of empathy that enabled recognition of autonomy within others. The combined result of the degradation of both human autonomy and empathy through digital means is the undermining of human dignity. As human dignity is undermined, human rights violations increase. This results in norms that regress away from expanded concepts of human dignity.

The regression of norms in digital spaces is remarkably progressed. Technical infrastructures are well on their way towards norm cascades if not already progressing towards internalization in discrete areas of data collection and usage. Science fiction is replete with the post norm cascade and internalization phases of the current regression of rights in digital spaces. One only need read Orwell's *1984, Ray Bradbury's Fahrenheit 451, Aldous Huxley's A Brave New World or more modern works by Cory Doctrow to gain a glimpse into the future where digital rights are consumed by advances in technologies. Ron Deibert is prescient in stating that a "Reset" is needed* [14]. Shoshana Zuboff's work serves as a canary in the proverbial coal mine warning that the world we are developing is not entirely as it seems [15]. There is a need to elevate a discourse of human rights in digital spaces. A solution has arisen in part in the European Union through the GDPR, but the networked nature of the Internet and the competing interests of states and their domestic Internets often forces rights considerations from the forefront to mere afterthoughts. It is unlikely that there will ever be a utopia predicated on rights-based discourses and norms, but neither should there be a dystopia. Recognizing and including digital rights concerns in broader human rights agendas is a critical component of countering regressive norms. Examining the impact of technologies on human autonomy and empathy and subsequently the conceptualization of human dignity of users is a critical first step in securing digital rights. Considering how to safeguard autonomy, empathy, and dignity in the development of new technologies is a critical first step to protecting digital human rights. This becomes increasingly important in an age when human rights are increasingly digital rights and digital rights are human rights.

References

[1] A. Wendt, "Anarchy is what states make of it," *International Organization*, vol. 46, no. 2, 1992.

[2] M. Finnemore, K. Sikkink, "International norm dynamics and political change," *International Organization*, vol. 52, no. 4, pp. 887–917, 2003, doi: 10.1162/002081898550789.

[3] D. Ronald, *Access contested: Security, identity, and resistance in Asian cyberspace information revolution and global politics*. Cambridge, MA: MIT Press, 2012.

[4] R. Deibert, J. Palfrey, R. Rohozinski, J. Zittrain, *Access controlled: The shaping of power, rights, and rule in cyberspace*. Cambridge, MA: MIT Press, 2010.

[5] R. Deibert, J. Palfrey, R. Rohozinski, J. Zittrain, Access denied: *The practice and policy of global internet filtering.* Cambridge, MA: MIT Press, 2008.

[6] A. Brantly, "The Cyber Losers," *Democracy and Security,* vol. 10, no. 2, pp. 132–155, 2014, doi: 10.1080/17419166.2014.890520.

[7] R. MacKinnon, *Consent of the networked: The worldwide struggle for Internet freedom*. New York, NY: Basic Books, 2012.

[8] L. Diamond, "Liberation Technology," *Journal of Democracy,* vol. 21, no. 3, pp. 69–83, 2010, doi: 10.1353/jod.0.0190.

[9] T. Zeynep, *Twitter and tear gas: The power and fragility of networked protest*. New Haven, CT: Yale University Press, 2017.

[10] A. Brantly, "From cyberspace to independence square: Understanding the impact of social media on physical protest mobilization during Ukraine's Euromaidan revolution," *Journal of Information Technology Politics*, pp. 1–19, 2019, doi: 10.1080/19331681.2019.1657047.

[11] A. Brantly. (2014, Jan. 24). *You were identified as a participant in a mass disturbance.* [Online]. Available: https://www.nditech.org/you-were-identified-participant-mass-disturbance. [Accessed: Nov. 29, 2022].

[12] B. Marczak, J. Scott-Railton. (2016, May 29). *Keep calm and (don't) enable macros: A new threat actor targets uae dissidents – the citizen lab.* [Online]. Available: https://citizenlab.ca/2016/05/stealth-falcon/. [Accessed: Nov. 29, 2022].

[13] J. Scott-Railton, S. Hardy. (2014, Dec. 18). *Malware attacks targeting syrian isis critic*s. [Online]. Available: http://citizenlab.ca/2014/12/malware-attack-targeting-syrian-isis-critics/. [Accessed: Nov. 29, 2022].

[14] R. Deibert, *Reset: Reclaiming the Internet for civil society*. Toronto, ON: Anansi, 2020.

[15] S. Zuboff, *The age of surveillance capitalism: the fight for a human future at the new frontier of power*, New York: PublicAffairs Press, 2019.

[16] H. Lynn, *Inventing human rights: A history*. New York: W.W. Norton & Company, 2007.

[17] R. Dean, *The value of humanity in Kant's moral theory*. New York: Oxford University Press, 2006.

[18] D. von der Pfordten, "On the dignity of man in Kant," *Philosophy*, vol. 84, no. 3, pp. 371–391, 2009, doi: 10.1017/s0031819109000370.

[19] M. Wollstonecraft, *A vindication of the rights of woman. Harmondsworth*. UK: Penguin Books, 1975.

[20] C. Beccaria, G. Newman, P. Marongiu, *On crimes and punishments*. New Brunswick, NJ: Transaction Publishers, 2009, doi: 10.4324/9781315125527.

[21] Voltaire, S. Harvey, *Treatise on tolerance*. Cambridge, UK: Cambridge University Press, 2000.

[22] E. Weitz, *A world divided: the global struggle for human rights in the age of nation-states.* Princeton, NJ: Princeton University Press, 2019.

[23] H. Arendt, *The origins of totalitarianism*. New York: Harcourt Brace, 1985.

[24] J. Kingdon, *Agendas, alternatives, and public policies*. New York: Longman, 2003.

[25] J. Donnelly, *Universal human rights in theory and practice*, 3rd ed. Ithaca, NY: Cornell University Press, 2013.

[26] United Nations, *Universal declaration of human rights.* [Online]. Available: https://www.un.org/en/aboutus/universal-declaration-of-human-rights. [Accessed: Mar. 9, 2022].

[27] J. Habermas, "The concept of human dignity and the realistic utopia of human rights," *Metaphilosophy*, vol. 41, no. 4, 2010.

[28] A. Moravcsik, "Explaining international human rights regimes: Liberal theory and Western Europe," *European Journal of International Relations*, vol. 1, no. 2, pp. 157–189, 1995.

[29] T. Solomon, "Norms and human rights in international relations," *Political Studies Review*, vol. 4, no. 1, pp. 36–47, 2005, doi: 10.1111/j.1478-9299.2006.00038.x.

[30] M. Caprioli, P. F. Trumbore, "Human rights rogues in interstate disputes, 1980–2001," *Journal of Peace Research*, vol. 43, no. 2, pp. 131–148, 2006, doi: 10.1177/0022343306061356.

[31] E. Neumayer, "Do international human rights treaties improve respect for human rights?," *Journal of Conflict Resolution*, vol. 49, no. 6, pp. 925–953, 2005, doi: 10.1177/0022002705281667.

[32] K. Sikkink, "Human rights, principled issue-networks, and sovereignty in Latin America," *International Organization*, vol. 47, no. 3, pp. 411–441, 1993, doi: 10.1017/s0020818300028010.

[33] B. Simmons, *Mobilizing for human rights: international law in domestic politics*. Cambridge, UK: Cambridge University Press, 2009.

[34] J. Abbate, *Inventing the Internet*. Cambridge, MA: MIT Press, 2000.

[35] K. Hafner, M. Lyon, *Where wizards stay up late: The origins of the Internet.* Simon & Schuster, 1996.

[36] P. Baran. (1962). *On distributed communications networks*. [Online]. Available: https://pages.cs.wisc.edu/~akella/CS740/F08/740-Papers/Bar64.pdf. [Accessed: Nov. 29, 2022].

[37] F. Heart. (1970). *Interface message processors for the ARPA computer network*. [Online]. Available: https://apps.dtic.mil/sti/pdfs/AD0709621.pdf. [Accessed: Nov. 29, 2022].

[38] V. G. Cerf, R. E. Kahn, "A protocol for packet network intercommunication," *Data Communications of the IEEE Communications Society*, p. 1–13, 1974.

[39] L. DeNardis, *Protocol politics: the globalization of Internet governance*. Cambridge, MA: MIT Press, 2009.

[40] A. Brantly, "A holistic approach to the encryption debate," in *Cyber insecurity: navigating the perils of the next information age*, T. Herr, R. Harrison, Eds. Lanham, Maryland: Rowman & Littlefield, 2016.

[41] D. Kehl, A. Wilson, K. Bankston. (2015). *Doomed to repeat history? Lessons from the Crypto Wars of the 1990s, Open Technology Institute.* [Online]. Available: https://static.newamerica.org/attachments/3407-doomed-to-repeat-history-lessons-from-the-crypto-wars-of-the-1990s/Crypto%20Wars_ReDo.7cb491837ac541709797bdf868d37f52.pdf. [Accessed: Nov. 29, 2022].

[42] A. F. Brantly, "Conceptualizing cyber policy through complexity theory," *Journal of Cyber Policy,* pp. 1–15, 2019, doi: 10.1080/23738871.2019.1583763.

[43] S. K. Pell, "You can't always get what you want: How will law enforcement get what it needs in a Post-CALEA, cybersecurity-centric encryption era?," *North Carolina Journal of Law and Technology*, vol. 17, no. 4, pp. 599–643, 2016.

[44] H. Klein, "ICANN and internet governance: leveraging technical coordination to realize global public policy," *Information Soc*, vol. 18, no. 3, pp. 193–207, 2011, doi: 10.1080/01972240290074959.

[45] J. P. Barlow. (1996). *A declaration of the independence of cyberspace, Electronic Frontier Foundation.* [Online]. Available: https://www.eff.org/cyberspace-independence. [Accessed: Aug. 30, 2021].

[46] M. Webb, C. Doctorow. (2020). *Coding democracy: how hackers are disrupting power, surveillance, and authoritarianism.* [Online]. Available: https://img1.od-cdn.com/ImageType-100/0111-1/{024C8725-D290-45E4-944F-0057B65CFB00}Img100.jpg. [Accessed: Nov. 29, 2022].

[47] J. Menn, *Cult of the Dead Cow: How the original hacking supergroup might just save the world.* New York: PublicAffairs, 2019.

[48] H. Abelson, R. Anderson. S.M. Bellovin, J. Benaloh, M. Blaze, W. Diffie, J. Gilmore, M. Green, S. Landau, P.G. Neumann, R.L. Rivest, J.I. Schiller, B. Schneier, M. Specter, D.J. Weitzner, "Keys under doormats: mandating insecurity by requiring government access to all data and communications," *Journal of Cybersecurity*, vol. 44, no. 1, p. tyv009–11, 2015, doi: 10.1093/cybsec/tyv009.

[49] W. Diffie, S. Landau, *Privacy on the line, updated and expanded edition*. Cambridge, MA: MIT Press, 2007.

[50] M. L. Mueller, *Networks and states: The global politics of internet governance.* Cambridge, MA: MIT Press, 2013.

[51] L. Denardis, *The global war for Internet governance.* New Haven, CT: Yale University Press, 2014.

[52] A. Greenberg, *This machine kills secrets: How WikiLeakers, cypherpunks and hacktivists aim to free the world's information*. New York: Dutton, 2012.

[53] J. Zittrain, *The future of the Internet and how to stop it.* New Haven, London: Yale University Press, 2008.

[54] C. Shirky, *Here comes everybody: The power of organizing without organizations.* New York: Penguin Press, 2008.

[55] H. R. Clinton. (2011, Feb. 15). *Remarks on Internet freedom*. [Online]. Available: https://www.eff.org/files/filenode/clinton_internet_rights_wrongs_20110215.pdf. [Accessed: Nov. 29, 2022].

[56] BBC News. (2014, Jan. 17). *Edward Snowden: Leaks that exposed US spy.* [Online]. Available: http://www.bbc.com/news/world-us-canada-23123964. [Accessed: Nov. 29, 2022].

[57] G. Greenwald. (2014). *No place to hide: Edward Snowden, the NSA, and the US surveillance state*. [Online]. Available: http://www.glenngreenwald.net/. [Accessed: Nov. 29, 2022].

[58] E. Morozov, *The net delusion: The dark side of internet freedom*. New York: Public Affairs, 2011.

[59] P. N. Howard, S. D. Agarwal, M. M. Hussain, "When do states disconnect their digital networks? Regime responses to the political uses of social media," *The Communication Review*, vol. 14, no. 3, pp. 216–232, 2011, doi: 10.1080/10714421.2011.597254.

[60] M. M. Hussain, P. N. Howard, "What best explains successful protest cascades? ICTs and the fuzzy causes of the arab spring," *International Studies Review*, vol. 15, no. 1, pp. 48–66, 2013, doi: 10.1111/misr.12020.

[61] Y. Benkler, *The wealth of networks: how social production transforms markets and freedom.* New Haven, CT: Yale University Press, 2006.

[62] J. Scott-Railton, B. Marczak, S. Anstis, B. A. Razzak, M. Crete-Nishihata, R. Deibert. (2018). *RECKLESS VI: Mexican journalists investigating cartels targeted with nso spyware following assassination of colleague, The Citizen Lab*. [Online]. Available: https://tspace.library.utoronto.ca/bitstream/1807/96737/1/Report%23116--Reckless%20VI.pdf. [Accessed: Nov. 29, 2022].

[63] B. Marczak, A. Abdulemam, N. Al-Jizawi, S. A. Berdan, J. Scott-Railton, R. Deibert. (2021, Aug. 24). *From Pearl to Pegasus Bahraini government hacks activists with NSO Group Zero-Click iPhone exploits, The Citizen Lab.* [Online]. Available: https://citizenlab.ca/2021/08/bahrain-hacks-activists-with-nso-group-zero-click-iphoneexploits/. [Accessed: Aug. 30, 2022].

[64] K. Zetter. (2021, July 22). *The NSO 'surveillance list': What it is and isn't, Zero Day*. [Online]. Available: https://zetter.substack.com/p/the-nso-surveillance-list-what-it [Accessed: Aug. 30, 2022].

[65] P. Tucker, *The naked future: What happens in a world that anticipates your every move.* New York: Current, 2014.

[66] S. J. Shackelford, *Internet of things.* New York: Oxford University Press, 2020.

[67] J. Lukito, "Coordinating a multi-platform disinformation campaign: Internet research agency activity on three US social media platforms, 2015 to 2017," *Political Communication*, vol. 37, no. 2, pp. 1–18, 2019, doi:10.1080/10584609.2019.1661889.

[68] P. N. Howard, M. M. Hussain, "The role of digital media," *Journal of Democracy*, vol. 22, no. 3, pp. 35–48, 2011, doi: 10.1353/jod.2011.0041.

[69] K. C. Desouza, A. Ahmad, H. Naseer, M. Sharma, "Weaponizing information systems for political disruption: The actor, lever, effects, and response taxonomy (ALERT)," *Computers and Security,* vol. 88, p. 101606, 2019, doi: 10.1016/j.cose.2019.101606.

[70] T. Lehtiniemi, "Personal data spaces: An intervention in surveillance capitalism?" *Surveillance & Society*, vol. 15, no. 5, pp. 626–639, 2017, doi: 10.24908/ss.v15i5.6424.

[71] M. Zajc, "The social media dispositive and monetization of user-generated content," *Information Society*, vol. 31, no. 1, pp. 61–67, 2014, doi: 10.1080/01972243.2015.977636.

[72] S. Brewster et al., "Social media as a passive sensor in longitudinal studies of human behavior and wellbeing," *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–8, 2019, doi: 10.1145/3290607.3299065.

[73] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, A. Pentland, "Once upon a crime: Towards crime prediction from demographics and mobile data," *Arxiv*, 2014, doi: 10.48550/arXiv.1409.2983.

[74] R. Bolla, F. Davoli, "Road traffic estimation from location tracking data in the mobile cellular network," vol. 3, pp. 1107–1112, 2000, doi: 10.1109/wcnc.2000.904783.

[75] D. Oosterlinck, D. F. Benoit, P. Baecke, N. V. de Weghe, "Bluetooth tracking of humans in an

indoor environment: An application to shopping mall visits," *Applied Geography*, vol. 78, pp. 55–65, 2017, doi: 10.1016/j.apgeog.2016.11.005.

[76] S. Abdelhamid, H. S. Hassanein, G. Takahara, "Vehicle as a mobile sensor," *Procedia Computer Science*, vol. 34, pp. 286–295, 2014, doi: 10.1016/j.procs.2014.07.025.

[77] S. Feldstein, *The rise of digital repression: How technology is reshaping power, politics, and resistance*. New York: Oxford University Press, 2021.

[78] T. M. Payton, T. Claypoole, *Privacy in the age of big data: recognizing threats, defending your rights, and protecting your family*. Lanham, MD: Rowman & Littlefield, 2014.

[79] A. Hern. (2018, Apr. 17). *Facebook admits tracking users and non-users off-site, The Guardian.* [Online]. Available: https://www.theguardian.com/technology/2018/apr/17/facebook-admits-tracking-users-and-nonusers-off-site. [Accessed: Mar. 14, 2022].

[80] X. Qiang, "The road to digital unfreedom: President Xi's surveillance state," *Journal of Democracy*, vol. 30, no. 1, pp. 53–67, 2019, doi: 10.1353/jod.2019.0004.

[81] C. Pope, "Biometric data collection in an unprotected world exploring the need for federal legislation," *Journal of Law and Policy*, vol. 26, no. 2, pp. 769–803, 2018.

[82] G. King, J. Pan, M. E. Roberts, "How censorship in China allows government criticism but silences collective expression," *American Political Science Review*, vol. 107, no. May, p. 326–343, 2012, doi: 10.1017/s0003055413000014.

[83] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, Y. Moreno, "The Dynamics of Protest Recruitment through an online network," *Scientific Reports*, vol. 1, no. 1, p. 197, 2011, doi: 10.1038/srep00197.

[84] K. Hylton, Y. Levy, L. P. Dringus, "Utilizing webcam-based proctoring to deter misconduct in online exams," *Computers & Education*, vol. 92, pp. 53–63, 2016, doi: 10.1016/j.compedu.2015.10.002.

[85] G. A. Teixeira, M. M. da Silva, R. Pereira, "The critical success factors of GDPR implementation: a systematic literature review," *Digital Policy Regulation and Governance*, vol. 21, no. 4, pp. 402–418, 2019, doi: 10.1108/dprg-01-2019-0007.

[86] W. L. Youmans, J. C. York, "Social media and the activist toolkit: User agreements, corporate interests, and the information infrastructure of modern social movements," *Journal of Communication*, vol. 62, no. 2, pp. 315–329, 2012, doi: 10.1111/j.1460-2466.2012.01636.x.

[87] E. P. Robinson, Y. Zhu, "Beyond 'I agree': Users' understanding of web site terms of service," *Social Media + Society*, vol. 6, no. 1, 2020, doi: 10.1177/2056305119897321.

[88] D. Hunter, N. Evans, "Facebook emotional contagion experiment controversy," *Research Ethics*, vol. 12, no. 1, pp. 2–3, 2016, doi: 10.1177/1747016115626341.

[89] P. Gerbaudo, F. Marogna, C. Alzetta, "When 'positive posting' attracts voters: User engagement and emotions in the 2017 UK election campaign on facebook," *Social Media + Society*, vol. 5, no. 4, 2019, doi: 10.1177/2056305119881695.

[90] A. G. Shu-Sha, H. Sophia, C. Jennifer, R. Andrea, "Social media use and empathy: A mini meta-analysis," *Social Networking*, vol. 8, no. 4, pp. 147–157, 2019, doi: 10.4236/sn.2019.84010.

[91] L. V. Bryant, "The youtube algorithm and the alt-right filter bubble," *Open Information Science*, vol. 4, no. 1, pp. 85–90, 2020, doi: 10.1515/opis-2020-0007.

[92] A. J. Flanagin, C. Flanagin, J. Flanagin, "Technical code and the social construction of the internet," *New Media & Society*, vol. 12, no. 2, pp. 179–196, 2010, doi: 10.1177/1461444809341391.

[93] S. U. Noble, "Algorithms of oppression: How search engines reinforce racism", *NYU Press*, pp. 119–133, 2018, doi: 10.2307/j.ctt1pwt9w5.8.

[94] M. Hildebrandt et al., "Mitigating bias in algorithmic hiring," Proceedings of the 2020 Conference on Fairness, Accountability and Transparency, pp. 469–481, 2020, doi: 10.1145/3351095.3372828.

[95] M. Oluwasanmi, "Algorithms and the border: The human rights implications of automated decision systems in Canadian immigration," *Federalism-E*, vol. 22, no. 1, 2021.

[96] J. Gerards, "The fundamental rights challenges of algorithms," *Netherlands Quarterly of Human Rights*, vol. 37, no. 3, pp. 205–209, 2019, doi: 10.1177/0924051919861773.

[97] L. McGregor, D. Murray, V. Ng, "International human rights law as a framework for algorithmic accountability," *International and Comparative Law Quarterly*, vol. 68, no. 2, pp. 309–343, 2019, doi: 10.1017/s0020589319000046.

[98] J. S. Hiller, G. Berger-Walliser, A. F. Brantly, "Critical protection for the network of persons," *Journal of Law and Social Change*, vol. 25, no. 2, pp. 117–152, 2021.

[99] C. W. Fitzgerald, A. F. Brantly, "Subverting reality: The role of propaganda in 21st century intelligence," *International Journal of Intelligence and Counterintelligence*, vol. 30, no. 2, pp. 215–240, 2017, doi: 10.1080/08850607.2017.1263528.

[100] G. Bolsover, P. Howard, "Computational propaganda and political big data: Moving toward a more critical research agenda," *Big Data*, vol. 5, no. 4, pp. 273–276, 2017, doi: 10.1089/big.2017.29024.cpr.

[101] P. N. Howard, S. Woolley, R. Calo, "Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration," *Journal of Information Technology and Politics*, vol. 15, no. 2, pp. 1–13, 2018, doi: 10.1080/19331681.2018.1448735.

[102] R. S. Baker, A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, 2021, doi: 10.1007/s40593-021-00285-9.

[103] T. Panch, H. Mattie, R. Atun, "Artificial intelligence and algorithmic bias: Implications for health systems," *Journal of Global Health*, vol. 9, no. 2, 2019, doi: 10.7189/jogh.09.020318.

[104] A. Brantly, N. D. Brantly, "Patient-centric cybersecurity," *Journal of Cyber Policy*, vol. 5, no. 3, pp. 1–20, 2020, doi: 10.1080/23738871.2020.1856902.

[105] A. Završnik, "Algorithmic justice: Algorithms and big data in criminal justice settings," *European Journal of Criminology*, vol. 18, no. 5, pp. 623–642, 2021, doi: 10.1177/1477370819876762.

[106] K. O'Hara and W. Hall, *Four internets: Data, geopolitics, and the governance of cyberspace.* New York: Oxford University Press, 2021.

[107] M. Milton, *Will the Internet fragment?: Sovereignty, globalization and cyberspace.* Cambridge: Polity Press, 2017.

[108] C. Baxter, O. Tkacheva, M. C. Libicki, L. H. Schwartz, J. E. Taylor, J. Martini, *Internet freedom and political space*. Santa Monica, CA: The RAND Corporation, 2013.

[109] M. E. Roberts, *Censored: Distraction and diversion inside China's great firewall.* Princeton, NJ: Princeton University Press, 2018.

[110] V. Carty, *Social movements and new technology.* New York and London: Routledge, 2015.

[111] M. Joyce, Ed., *Digital activism decoded: The new mechanics of change.* New York: International Debate Education Association, 2010.

[112] B. Rolfe, "Building an electronic repertoire of contention," *Social Movement Studies,* vol. 4, no. 1, pp. 65–74, 2005, doi: 10.1080/14742830500051945.

[113] R. Rohrschneider, R. Dalton, "A global network? Transnational cooperation among environmental groups," *The Journal of Politics*, vol. 64, no. 2, pp. 510–533, 2002.

[114] S. H. Kamel, "Egypt's ongoing uprising and the role of social media: Is there development?" *Information Technology for Development*, vol. 20, no. 1, p. 78–91, 2014, doi: 10.1080/02681102.2013.840948.

[115] A. Karatzogianni, *Cyber conflict and global politics*. Abingdon, Oxon: Routledge, 2009, doi: 10.4324/9780203890769.

[116] V. P. Shannon, "Norms are what states make of them: The political psychology of norm violation," *International Studies Quarterly*, vol. 44, pp. 293–316, 2000.

[117] S. Moyn, *Not enough human rights in an unequal world.* Cambridge, MA: Harvard University Press, 2018.

[118] N. Nanjala, *Digital democracy, analogue politics: How the internet era is transforming Kenya.* London, UK: Zed Books LTD, 2018.

The Cybersecurity Obligations of States Perceived as Platforms: Are Current European National Cybersecurity Strategies Enough?

Vagelis Papakonstantinou | **Faculty of Law and Criminology, Vrije Universiteit Brussel, Belgium**, ORCID: 0000-0002-2536-2951

Abstract

Cybersecurity is a relatively recent addition to the list of preoccupations for modern states. The forceful emergence of the internet and computer networks and their subsequent prevalence quickly brought this to the fore. By now, it is inconceivable that modern administrations, whether public or private, can exist entirely outside the digital realm. Nevertheless, with great opportunities also comes great risk. Attacks against computer systems quickly evolved from marginalised incidents to matters of state concern. The exponential increase in the importance of cybersecurity over the past few years has led to a multi-level response. New policies, followed by relevant laws and regulations, have been introduced at national and international levels. While modern states have therefore been compelled to devise concrete cybersecurity strategies in response to potential threats, the most notable aspect of these strategies is their resemblance to one another. Such uniform thinking could develop into a risk *per se*: challenges may appear unexpectedly, given the dynamic nature of the internet and the multitude of actors and sources of risk, which could put common knowledge, or what may be called conventional wisdom, to the test at a stage where the scope for response is limited. This paper builds upon the idea of national states being perceived as platforms within the contemporary digital and regulatory environment. Platforms are in this context information structures or systems, whereby the primary role of states acting as platforms is that of an information broker for its citizens or subjects.

Corresponding author:
Vagelis Papakonstantinou,
Faculty of Law and Criminology,
Vrije Universiteit Brussel,
Pleinlaan 2, 1050
Brussels, Belgium, ORCID:
0000-0002-2536-2951;
vagelis.papakonstantinou@vub.be

This role takes precedence even over the fundamental obligation of states to provide security; it calls upon them first to co-create (basic) personal data, and then to safely store and further transmit such data. Once the key concept of states as platforms has been elaborated in section 2, this paper then presents the concrete consequences of this approach within the cybersecurity field. In section 3, former off-line practices for safely storing personal information, undertaken by states within their role as platforms, are contrasted with the challenges posed by the digitisation of information. The focus is then turned in section 4 to the EU, and the NIS Directive's obligation upon Member States to introduce and implement national cybersecurity strategies, which are therefore examined under the lens introduced in section 2. Finally, specific points for improvement and relevant recommendations for these cybersecurity strategies are presented in section 5.

Keywords

data localisation, digital sovereignty, national cybersecurity strategies, states as platforms

Cite this article as: V. Papakonstantinou, "The Cybersecurity Obligations of States Perceived as Platforms: Are Current European National Cybersecurity Strategies enough?," ACIG, vol. 1, no. 1, pp. 56–68, 2022, DOI: 10.5604/01.3001.0016.1237

1. Introduction

Cybersecurity is a relatively recent addition to the list of preoccupations for modern states. The forceful emergence of the internet and computer networks and their subsequent prevalence quickly brought this to the fore. The use of the internet in public and private administrations developed rapidly from a useful accessory into an inherent, embedded element of all relevant policies and strategies. By now, it is inconceivable that modern administrations, whether public or private, can exist entirely outside the digital realm. Nevertheless, with great opportunities also comes great risk. Attacks against computer systems quickly evolved from marginalised incidents to matters of state concern. Cybersecurity, notwithstanding the definition found in the EU's Cybersecurity Act[1], is a broad term encompassing anything from private security on a standalone computer for personal use to state security and cyberwarfare. It is under this latter context that the term will be used in this paper, to refer to the obligation of modern states to provide and prioritise a secure cyber environment for their citizens or subjects, in order to protect them against cyberthreats and cyberattacks.

The exponential increase in the importance of cybersecurity over the past few years has led to a multi-level response. New policies, followed by relevant laws and regulations, have been introduced at national and international level. New academic interest has emerged (as is apparent from the release of this first issue of an aspiring new academic journal), adding to the traditional studies on security. A new market has also emerged, aimed at satisfying increased consumer and business needs in

1 —— "Cybersecurity means the activities necessary to protect network and information systems, the users of such systems, and other persons affected by cyber threats", Art. 2(1), Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) no 526/2013 (Cybersecurity Act).

this sphere. Perhaps the most notable aspect of these responses, however, is that modern states have been compelled to devise concrete cybersecurity strategies. The aim of these strategies is twofold, both to protect and to reassure. States need to protect their citizens and assets from cyberattacks and cyberthreats through specific cybersecurity measures[2]. They also need to be able to demonstrate to their citizens that they are aware of the risk and are taking mitigating measures in tandem that respect the political, historical and cultural circumstances of the societies concerned. National cybersecurity strategies, as published on the internet, need to attain both of these targets.

Nevertheless, the commonality of computer network technologies and of the internet itself (including the digital means to cause harm) has led to considerable harmonisation among state responses. In other words, national cybersecurity strategies mandated by EU law (as made available over the internet) more or less resemble one another. Admittedly, harmonisation has been an explicit aim at both the regional (EU) and international level. Either under a formal legal obligation or within a best practice context, modern states have formulated public national cybersecurity strategies that are similar, both in terms of their assumptions and with regard to their aims and purposes (typically also including the means to accomplish them). However, such uniform thinking could develop into a risk *per se*: challenges, in the form of cyber risks, may arise unexpectedly, given the dynamic nature of the internet and the multitude of actors and sources of risk, which could put common knowledge (or received wisdom) to the test at a stage where the scope for response is constrained.

This paper builds upon the idea of national states being perceived as platforms within the contemporary digital and regulatory environment [1]. In this context, it elaborates upon a concrete consequence of states-as-platforms within the cybersecurity field. To this end, section 2 introduces the idea of states as platforms; section 3 then particularizes this general discussion by specifically referring to the cybersecurity field, in an attempt to highlight specific consequences of states being perceived as platforms. The focus is then turned to the EU, and the NIS Directive's obligation upon Member States to introduce and implement national cybersecurity strategies, which are therefore examined under the lens introduced in section 2. Finally, specific points for improvement in these cybersecurity strategies are presented in section 5.

2. States as Platforms

The recent adoption, in July 2022, by the European Parliament of the Digital Services Act[3] means that, once it officially comes into effect, it will formally introduce into EU law the term "online platforms": these (at least according to the Commission's original proposal) are meant to constitute "a provider of a hosting service which, at the request of a recipient of the service, stores and disseminates to the public information"[4], whereby a hosting service, in turn, "consists of the storage of information provided by, and at the request of, a recipient of the service"[5]. Therefore, between the DSA and the Digital Markets Act[6], which was simultaneously adopted as part

2 —— For Europe, this obligation is introduced most prominently in the text of the NIS Directive (Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union), whereby Member States are obliged to adopt a national strategy on the security of network and information systems (Art. 1 par. 2(a), where the latter is defined as "the ability of network and information systems to resist, at a given level of confidence, any action that compromises the availability, authenticity, integrity or confidentiality of stored or transmitted or processed data or the related services offered by, or accessible via, those network and information systems" in its Art. 2(2).

3 —— Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM/2020/825 final (the "DSA").

4 —— Art. 2, point (h), DSA.

5 —— Art. 2, point (f), DSA.

6 —— Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act), COM/2020/842 final (the "DMA").

of a single "Digital Services Act package", a comprehensive framework for the regulation of online platforms is introduced in EU law, the first of its kind both in Europe and internationally.

What constitutes a platform, the projection of which in the digital environment has recently attracted so much of the EU legislator's attention? The legislative definition seen above focuses on what online platforms do, not on what they are. In essence, online platforms are information structures or systems based on software. It is in this (digitised) context that the term is used in EU law. However, in the real, non-digital world the term literally denotes a "flat raised area or structure" (Cambridge dictionary) or "a raised level surface on which people or things can stand" (Oxford dictionary). In real-world usage the term has been employed metaphorically to denote sets of policies or ideas. What is common in both cases is differentiation, even exceptionalism. The concept of a platform may represent something raised above the mundane, or even a singular grouping of ideas, which differentiates it from all others. But there is also the matter of context – where a platform has interconnectedness with other platforms around it: a platform cannot exist in the void. Finally, platforms are structured around basic rules (whether behavioural, regulatory, or other) that are common to all their supporters or users. It is perhaps these characteristics of real-world platforms that rendered the word apt for describing large information systems in the digital world.

The EU's first attempt to regulate online platforms came through the so-called P2B Regulation (platforms-to-business relations Regulation)[7]. In the Commission's words, the P2B Regulation is the "first ever set of rules for creating a fair, transparent and predictable business environment for smaller businesses and traders on online platforms" [2]. While the P2B Regulation therefore aims at regulating the relationship of online platforms with their business users, it is the DSA, and to a lesser extent the DMA, that are expected to govern the other side of the spectrum, namely the relationships between online platforms and their individual users or consumers.

However, European regulatory innovation in the field perhaps invites a different viewpoint: could states themselves be considered as platforms? What if this newly finalised EU regulatory framework was applied to states also? What insights could be derived into the role of states from EU online platform regulation?

In order to address these questions, the first step lies in uncovering the basic role of states as information brokers. Although this realisation did not become evident until the Information Revolution gave importance to the role of information in human lives, states – within the meaning of organised societies – are first and foremost information brokers for their subjects or citizens[8]. At the moment of birth humans are vested with state-provided information: a name[9], as well as a specific nationality [6, p. 75]. Without these a person cannot officially exist. A nameless individual is unthinkable in human societies. Although it is the family that provides a person with a name when he or she is born, without a specific mechanism to formally acknowledge it such a name could function only among a very small number of people[10]. It is therefore a state (in the above meaning) that, at first, validates a name for a person and then is responsible for its safekeeping

7 —— Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services.

8 —— Although the role of states assembling "informational capital" has been identified for example by Bourdieu [3, p. 213], this is different than the state's effort to "measure, count, assess, investigate".

9 —— Although this is certainly true in modern, centralised bureaucratic states, the same has arguably been the case in any organised society, regardless of whether it was within an iron age empire, a city-state, the Roman empire, Medieval Europe etc. **[4]**. In other words, ever since t he first organised human societies emerged, individuals needed to be registered, if not for anything else then for taxation and military service purposes [5, p. XI].

10 —— Although Herzog notes that "for many years historians assumed that there were absolutely no rules indicating who would be called what, or guaranteeing that a person would use the same name throughout his or her life" [7, p. 199], for the purposes of this paper actual use is irrelevant; keeping also in mind the state's best interest in having consistency of names, whether any given person lives an extremely static life and therefore is in no need of a formal name is beside the point.

through specific bureaucratic mechanisms (or at least its safekeeping is in the state's best interest). The second type of essential information provided by the state at the time of birth of any individual is nationality, in the sense of belonging to a specific state or organised society. Just as is true for names, a stateless person is unthinkable in human societies.

The above two sets of information are subsequently much more enriched within modern, bureaucratic states. Education and employment, family status, property rights, taxation and social security is all information (co-) created[11] by states and their citizens or subjects. For the purposes of this analysis, this type of personal information shall be designated as "basic personal data". It is after basic personal data have been created that the second, equally crucial, part of the role of states as information brokers comes into play: states safely store and further disseminate personal data. This is of paramount, fundamental importance to individuals. In order to go about their lives in any meaningful manner, it is imperative that individuals first have their basic personal data stored safely, and then for such data to be readily communicable by their respective state as required. As regards storage, individuals need their basic personal data stored securely for the duration of their lives and for a short period thereafter (at least until all their property rights expire). They need this information to be persistent and not to be tampered with, in order for them to be able to enter into any transaction with third parties over the course of their lives. Second, individuals need this information disseminated to third parties through the intermediation of the state granting validity to the transmission. Trust in human transactions is tacitly provided by the state, through its validation (or even direct transmission) of the personal information concerned.

Information brokerage is therefore the primary role of the state, which takes precedence over any other. No political or state organisation theory can provide individuals with any meaningful life, without their basic personal information safely stored and further transmittable. Accordingly, if a state "loses" a birth certificate or a family record, the persons concerned need to immediately replace them with the assistance of another state, otherwise they will be placed in a state of limbo – and thus in great insecurity. Ultimately, what has already been identified in Hobbes' *Leviathan* as the most fundamental role of a state, the provision of security, would be meaningless unless that state's function as an information broker has already occurred, meaning that the state knows who it has to protect[12].

Once the extremely important role of information brokerage for their citizens has been acknowledged, the next step for states is to relate this role with the platforms that have recently captured the EU legislator's attention. Can states in fact be viewed as platforms? Firstly, one could easily remove the digital elements in the EU's definition of online platforms. In essence, the DSA's definition may well apply in the real world too: platforms store and disseminate information to the public at the request of their users. A state viewed as a platform would then form the intermediary in an information flow from its citizens (users, individuals) to everybody else. From this perspective, platforms essentially coincide with the state as information broker, in the manner described above. Or, in other words, states have actually functioned as platforms, albeit in the real world, since the first organised societies emerged.

11 —— The role of the state, however, is not that of a trusted third party. The state does not simply safeguard information on its subjects that was created by the subjects themselves but instead actively participates in its creation, by establishing and maintaining the institutions within which creation of this information becomes possible.

12 —— Similarly, if under a different political theory the primary role of the state is justice, the state still needs to know who its recipients are.

3. Cybersecurity obligations of States-as-Platforms

The importance of such safekeeping cannot be overstated. As explained in section 2, a nameless or stateless person is out of the question. Similarly, within contemporary societies a person without any family, education or employment data risks living a marginalised and precarious existence. The only entities capable and in charge of safekeeping this information are states. It is their responsibility first to enable the creation and access of this information to individuals (the ownership question over such information notwithstanding), and, once that task has been completed, to make sure that this information remains available and transmittable to any third party at the request of the individuals concerned. States-as-platforms have exclusivity over this extremely important role.

If seen as platforms whose most fundamental role is to store and transmit the basic personal data of citizens to third parties, states carry a specific set of responsibilities. Within the traditional field of security the focus is on individuals themselves [8, 9]: their security, in the sense of physical and psychological well-being, as well as (depending on the theory adopted) their ability to flourish within society, largely dominate the relevant discourse. That same individual's information has attracted much less attention in this regard. However, from a states-as-platforms perspective, the focus turns clearly towards the basic personal data itself: if information brokerage is what states primarily do, and states are the primary providers of a right to security as their most basic *raison d'être*, then such information needs to remain secure, first and foremost.

The security of individuals' basic personal data is therefore crucial. The type of security measures assumed so far by states in order to warrant this inevitably stemmed from the nature of the data stored. Until recently, all basic personal data were registered in paper records, which were kept manually [10]. Digitisation of information came quite late in human history, and much later in public administration[13]. A number of important realisations arise from this understanding. The first is that paper record-creation and record-keeping remained one of the basic functions of state administrations. Once printing became available, paper public records were meticulously created by hand and organised in elaborate filing systems [11]. Their maintenance was of utmost importance: paper records necessary to carry out transactions (i.e. pertaining to living individuals) were carefully kept, updated and preserved. Photocopies and photography assisted this process. Although a relevant analysis of this process goes beyond the scope of this paper, here it is sufficient to note that paper record-keeping has been the norm until very recently, when it comes to states operating as information brokers for their citizens or subjects.

The second realisation refers to the fact that states kept all their basic personal data locally, meaning within their respective territories and jurisdictions. This was unavoidable, given the nature of the data concerned: paper records (or, much less, records kept in stone or any other material) could not be moved from one state to another, either for safekeeping or for any other conceivable reason. This is an important clarification, directly connected to contemporary discussions on digital sovereignty (see the analysis in section 4). Throughout their history (and under whatever political

13 —— The reason behind the GDPR's automated and non-automated files is a legacy provision of its predecessor, the 1994 Directive, which in turn included it because most public sector files in Europe were not digitized until the early nineties.

system they maintained) states were never locally separated from the basic personal data of their citizens or subjects. All creation, safekeeping and transmission was performed locally, within their borders; whenever a formal, case-specific transmission to a third state became necessary, elaborate schemes[14] were agreed among states, basically affording cross-platform transmission of information (in essence, platform interoperability).

The third realisation refers to the proportionality of the security measures assumed by states to protect the basic personal data of their citizens. Paper records containing their citizens' basic personal data were kept in state buildings. State buildings are protected both by state authorities (the police) and by the law: in most countries the destruction of public property is a serious crime. Of course, at a higher level, states themselves, including their paper records, are protected by their military forces against attacks from any other state. Consequently, as physical objects of great importance, public records (including the basic personal data) were kept with utmost care by states, raising the level of proportional protection measures to the highest level possible afforded by the state concerned.

Notwithstanding public record-keeping theory and practice, for the purposes of this paper it should be clarified that, as long as basic digital personal data are concerned, the obligations of states as platforms fall within the cybersecurity field. Once digitised, basic personal data lose their tangible nature and existence in the real world and exist only electronically (the same, of course, applying to digital-born data as well). They are no longer paper records, that need to be physically preserved and delivered manually upon request to any party concerned. They are digital records that are transmitted electronically, over the internet or otherwise. In addition, digital state records no longer have a real-world equivalent: electronically created state records containing basic personal data are not additionally printed in paper form, either for safekeeping or for any other purpose. Consequently, if they are deleted for any reason, electronic records are lost forever. While of course this has always been a risk with paper records too, which can be lost or destroyed through natural disasters (e.g. fire or flood) or wilful acts (e.g. war), the risk in such case is admittedly much lower: when it comes to electronic files, pressing a button may lead to the deletion of huge volumes of data in a split second, whereas the burning or flooding of paper records held by the state doesn't occur instantly, and the rate of destruction would most likely be limited due to localisation of the records concerned. The digitalisation of information has allowed state administrations to change their centuries-old methodology of record-creation and safekeeping, moving from a tangible to an intangible format. It is under this change that the states-as-platforms obligations within the cybersecurity context become visible.

A further realisation stems from the digitalisation of basic personal data: other than their increased transferability and, perhaps, vulnerability when compared to their (paper) predecessors, they also enable more efficient state administration. This is a realistic (electronic records are easier to manage than paper records by the same civil servant) and also powerful assumption that has had multiple legal repercussions in the past: namely, it led to the introduction of a new field of law – data protection law – in the 1970s [12, p. 50], and has also led to important case law, such as *Google Spain* and the right to de-listing from online search engines[15].

14 — See, for example, the Hague Treaty (Convention of 5 October 1961 Abolishing the Requirement of Legalisation for Foreign Public Documents (HCCH 1961 Apostille Convention)).

15 — The ruling of the CJEU specifically refers to "processing of personal data, such as that at issue in the main proceedings, carried out by the operator of a search engine is liable to affect significantly the fundamental rights to privacy and to the protection of personal data when the search by means of that engine is carried out on the basis of an individual's name, since that processing enables any internet user to obtain through the list of results a structured overview of the information relating to that individual that can be found on the internet — information which potentially concerns a vast number of aspects of his private life and which, without the search engine, could not have been interconnected or could have been only with great difficulty — and thereby to establish a more or less detailed profile of him. Furthermore, the effect of the interference with those rights of the data subject is heightened on account of the important role played by the internet and search engines in modern society, which render the information contained in such a list of results ubiquitous", Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González, Case C-131/12, p. 80.

Consequently, the automation of the processing of basic personal data is of great importance. While in peaceful Western societies this realisation is mostly a benign one, inviting analyses, for example, on how to balance data management optimisation against protection of individual rights, the future ought not be taken for granted: in the event of war, a foreign administration seizing the digitised state records of the defeated state's subjects and citizens will find in its hands a powerful tool of occupation and repression.

The above realisations are by no means intended to constitute an exhaustive analysis of the risks posed to modern states by the digitalisation of state records, particularly those including basic personal data. The aim of the analysis is to highlight the new challenges that states-as-platforms are faced with in the digital realm. While some of these challenges were also present in the past, under the basic role of states as information brokers as seen in section 2, they were largely tacitly mitigated, if not supressed, by the nature of the information *per se*: paper records are neither movable nor easily perishable or easily manageable. Digitised records, however, present none of these faculties: on the contrary, they are easily transmittable, deletable *en masse*, and automatically processable. It is precisely from this perspective that states-as-platforms need to take note, as part of their cybersecurity policies and strategies.

4. Two important shortcomings in European national cybersecurity strategies

The obligation of EU Member States to introduce national cybersecurity strategies ("NCSS") was formally introduced relatively late, in 2016, by the NIS Directive [13, p. 6]. Although cybersecurity risks were acknowledged at the EU level many years ago [14], and in spite of the fact that at the time when the NIS Directive came into effect a number of European countries had already introduced cybersecurity strategies within their respective jurisdictions [15, p. 7, 16, p. 55], horizontal implementation throughout Europe was achieved only through the NIS Directive. In addition to this basic contribution to Member States' cybersecurity, the NIS Directive's other major contribution was the delineation of the contents of such a strategy within its text: according to Article 7 par. 1, such a strategy would have to at least address seven topics, namely: (I) the objectives and priorities of the national strategy on the security of network and information systems, (II) a governance framework to achieve them, (III) the identification of measures relating to preparedness, response and recovery, (IV) an indication of the education, awareness-raising and training programmes relating to the national strategy on the security of network and information systems, (V) an indication of the research and development plans relating to the national strategy on the security of network and information systems, (VI) a risk assessment plan, and (VII) a list of the various actors involved in the implementation of such national strategy. The Commission's approach was later confirmed in the text of the EU Cybersecurity Act[16], as well as in the text of the NIS2 Directive[17]. The new EU's new cybersecurity strategy, released in late 2020, further articulated three areas of EU action, namely (a) resilience, technological sovereignty and leadership, (b) operational

16 — See its Art. 2(3).

17 — See Art. 7, Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive).

capacity to prevent, deter and respond, and (c) cooperation to advance a global and open cyberspace [17, p. 4].

In case they need it, European states are authorised by the NIS2 Directive to ask ENISA, the EU agency for cybersecurity for assistance[18]. On its part ENISA duly obliged, releasing a series of documents to this end, in view of the fact that its involvement in the field had started as early as 2012 [18]: among other things, ENISA has issued a relevant Good Practice Guide [19] and a National Capabilities Assessment Framework [20], together with a guide outlining good practices in innovation on cybersecurity under the NCSS [21]. ENISA also provides an online monitoring tool, the ENISA NCSS Map, conveniently listing all NCSS applicable in the EU, including their strategic objectives and model examples of implementation[19].

Although a detailed comparison of all EU Member States' NCSS lies outside the purposes of this paper, comparative reading immediately makes apparent the impression made upon them both by the European Commission's approach, as included in the NIS2 Directive, as well as by the ENISA guidance. Specifically, in observance to their NIS2 Directive obligations, all Member States have published NCSS, which in greater or lesser detail address the seven points of Article 7 seen above. In addition, all these NCSS have to a large extent taken into consideration ENISA's Good Practice Guide [22, p. 110], as acknowledged in ENISA's National Capabilities Assessment Framework: ".. the disparity between the different Member States makes it difficult to identify common activities or action plans among different national contexts, legal frameworks and political agendas. However, Member States' NCSS's often have strategic objectives articulated around the same topics. Thus, based on ENISA's previous work and the analysis of Member States' NCSS's, 22 strategic objectives were identified" [20, p. 11]. In this manner the NIS Directive seems to have attained a harmonisation effect, in the sense that a few years after its introduction all EU Member States' NCSS appear aligned[20]. Evidently, not all information pertaining to these NCSS is public: because this is ultimately a matter of national security, it is possible that Member States make public through their NCSS[21] only those parts of their actual cybersecurity strategies that are necessary to comply with EU law, not wishing to compromise in this manner any national security state secrets.

At any event, from a states-as-platforms perspective at least two shortcomings may be identified in current European NCSS implementations: first and foremost, they consider all information to be of equal status. Nevertheless, critical infrastructures notwithstanding, not all information is of equal importance to others. While long debates could be held within a risk assessment analysis over which data relating to critical infrastructure are more important than others, or which *digital assets* are of higher value, the fact remains that basic personal data ultimately trump all others: if names or nationality information are tampered with or even permanently deleted[22], the effect could be devastating for the individuals and the states concerned. The same would apply to family, education, employment, and tax data. From the point of view of individuals and states, any deletion or tampering with these data would have a devastating effect, whereas any unauthorised access to, for example, bank transactions or the transport system would of course create major problems but scarcely on a similar scale.

18 — See Art. 7 par. 4 of th e NIS2 Directive.

19 — Information from ENISA's website, https://www.enisa.europa.eu/topics/national-cyber-security-strategies/ncss-map.

20 — According to Baezner and Cordey, who carried out a comparative NCSS analysis n 2019, cybersecurity strategies shared "a number of common conceptual elements", differences being mostly traced "in where cybersecurity is positioned within the context of government structures, and who bears which responsibilities" [15, p. 4].

21 — And, in fact, translated into English for most countries as early as in 2013 [23, p. 7].

22 — It should be noted that this is a fundamentally different case to identity theft: while identity theft customarily pertains to fraud or other (cyber)crime, in this case state records containing basic personal data may be endangered for national security aims and purposes.

The second shortcoming refers to treatment of data security in the event of defeat. Admittedly, the case of loss is addressed through resilience in cybersecurity strategies: the ability of a state to recover in case its protective measures fail [24, p. 29, 25, p. 6]. Defeat is, however, something larger: It means that another state has assumed, through an aggressive act of war, the defeated party's role. What happens then? The reply invites ethical and technical considerations. Should the defeated state accept defeat and assist the winner in assuming its role of managing the lives of its newly acquired subjects, or not? Should state records, particularly including basic personal information, be seamlessly handed over, or not? Depending on the reply to these fundamental questions, different strategies need to be devised. Although addressing these questions would largely be dependent upon political, societal and financial factors, the fact remains that under a states-as-platforms perspective the states concerned need to have made up their mind on these topics and apply specific measures in their national cybersecurity strategies accordingly.

5. Points for improvement

A divergence between the obligations of states-as-platforms and current national cybersecurity strategies is therefore evident from the points made above. As seen in section 3, states viewed as information brokers carry increased responsibilities towards their citizens in view of the digitisation of information. Such increased responsibilities need to be reflected in their respective cybersecurity strategies. EU and Member State national cybersecurity strategies do not appear to fare well under the above criteria: because they are focused more on protective and mitigation measures, they do not take into account the sensitivity of certain categories of information or the event of failure to protect data or even suffer defeat in case of war. It is in this context that certain recommendations will be outlined in this section. This is done not with the intention of compiling a comprehensive list of cybersecurity measures to address the above concerns, but rather by way of presenting examples, in order to attempt a paradigm shift in contemporary cybersecurity national strategies' thinking.

From a cybersecurity perspective, states have to take into account the importance of basic personal information on the one hand, and their role as information brokers on the other hand. The Information Revolution only served to accentuate and bring to the fore their role and responsibility within the states-as-platforms context. Since being a nameless and stateless person is unthinkable in the modern age, states need to ensure that nothing happens to digital records in this regard. Records kept on paper benefited from natural protection, being hard to destroy completely (and even harder to alter) and kept in state buildings, protected by law, the police, and ultimately the military forces of the state concerned. On the contrary, state records that are either digitised or born-digital are easier to destroy or alter and may not even be stored in state-run premises but rather outsourced to the private sector, even outside state borders. Within a states-as-platforms context, all of the above factors need to change: born-digital or digitised state data, including a person's basic personal

data, need to acquire the highest level possible of protection as a digital record, must be kept by the state itself (not outsourced to the private sector), maintained within that state's borders, and ultimately protected physically and electronically by that state's military. It is only in this way that the state will be able to continue serving its fundamental role as an information broker to its citizens. Notwithstanding the adage that "100% security is impossible", the fact remains that states, as is true for records kept in stone or paper throughout human history, have to do everything within their power to keep basic personal data safe.

Once security has been provided, mitigation measures towards worst-case scenarios ought not be overlooked. For example, a successful cyberattack could achieve deletion or alteration of born-digital state records including basic personal data, therefore creating insurmountable difficulties to the state and individuals concerned: mitigation measures need not only be technical and organisational (for example, encryption of the data or their dispersal to several physical locations) but also legal, ultimately leading to proof of identity by real-world means. Similarly, in the event of war won by an aggressor state it must be assumed that all state records of the defeated party will be taken over as well. Because these records include the basic personal data of that defeated state's citizens, they could constitute an extremely powerful tool for oppression, mismanagement, reshaping the previous state's nation-building narrative or discouraging opposition. Individuals subject to digitised or born-digital records will have limited means of resistance available to them, in the sense of providing adequate proof to amend or restore their compromised state records. Mitigation measures within national cybersecurity strategies need to be employed in this regard.

Finally, on a less basic but also important level, states-as-platforms need to carefully and diligently preserve the digital footprint of their citizens as well. While this is of course an already acknowledged task, in most cases it is carried out as part of states' archivist or cultural heritage tasks[23]. However, within the context discussed above, digital preservation is no longer a cultural priority but also a security one. In the event of loss or alteration of state records, the digital lives of their citizens, even if created under an informal, i.e. private capacity, may serve as means of proof or digital evidence. They may serve to contradict affected state records or to cross-reference information in order to prove a claim that, after in the wake of a successful cyberattack, may no longer be tenable. As a result, states operating within a states-as-platforms context need to make the relevant provisions in their national cybersecurity strategies.

23 —— For ease of reference simply refer to the UK's National Archives webpage, where it is stated that "We are [...] the official archive and publisher for the UK Government, and for England and Wales. We are the guardians of over 1,000 years of iconic national documents. We are expert advisers in information and records management and are a cultural, academic and heritage institution. We fulfil a leadership role for the archive sector and work to secure the future of physical and digital records", as well as "We collect and secure the future of the government record, from Shakespeare's will to tweets from Downing Street, to preserve it for generations to come".

6. Conclusions

In 1669 the Venetians, leaving the island of Crete to its new occupiers, the Ottomans, negotiated and successfully managed to take the state archives with them to Italy. In the back of their minds they thought to re-establish themselves on the island in the future (something that they subsequently tried but failed to accomplish), and these records would be crucial in this regard [26, p. 203][24]. State records, particularly when including basic

24 —— The same negotiation seems to have occurred hundreds of years later, during the Greek and Turkish population "exchange" in 1924 [27, p. 324].

personal data, have long since been invaluable in the event of military conflict. Whether digitised or born-digital, such records set new standards and pose new challenges to this much older discussion.

Even though states have always operated first and foremost as information brokers for their citizens or subjects, it is the Information Revolution that has undeniably brought this role to the fore. Within a states-as-platforms context, states have increased responsibilities and obligations as regards their citizens' personal information, especially when referring to basic personal data. Questions of state survival and continuity, especially when placed alongside human survival and well-being in the case of war (or even defeat), need to be re-visited and re-assessed within the digital environment, where, among other things, digitised or born-digital state records are, by their very nature, easier to destroy, alter or transfer than their older paper counterparts.

It is from this, perhaps novel, approach that EU Member States' cybersecurity strategies also need to be assessed. In their current format (as far as openly made public), they suffer from at least two shortcomings when viewed from a states-as-platforms perspective: they treat all information as equal (even when taking into account the critical infrastructure discussion) and they take no account of the case of defeat. While this paper does not purport to compile a comprehensive list of mitigation measures in this regard, it makes the point for data localisation and exclusivity of state protection in order for states to adequately support their role as information platforms for their citizens [28].

References

[1] V. Papakonstantinou, "States as platforms following the new EU regulations on online platforms," *European View*, vol. 21, no. 2, pp. 214–222, 2022, doi: 10.1177/17816858221134748.

[2] European Commission. (2019, Feb. 14). *Press Release: Digital Single Market: EU negotiators agree to set up new European rules to improve fairness of online platforms' trading practices.* [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1168. [Accessed: Sep. 1, 2022].

[3] P. Bourdieu, *On the state: Lectures at the Collège de France, 1989–1992.* Cambridge: Polity, 2015.

[4] K. Breckenridge, S. Szreter, Eds., *Registration and recognition: Documenting the person in world history.* Oxford: Oxford University Press, 2012.

[5] C. A. Bayly, "Foreword," in *Registration and recognition: Documenting the person in world history*, K. Breckenridge, S. Szreter, Eds. Oxford: Oxford University Press, 2012.

[6] G. W. F. Hegel, *Hegel: Elements of the philosophy of right.* Cambridge: Cambridge University Press, 1991.

[7] T. Herzog, "Naming, identifying and authorizing movement in early modern Spain and Spanish America," in *Registration and recognition: Documenting the person in world history*, K. Breckenridge, S. Szreter, Eds. Oxford University Press, 2012.

[8] L. Lazarus, "Mapping the right to security," in *Security and human rights,* B. J. Goold, L. Lazarus, Eds. Oxford: Hart Publishing, 2007.

[9] S. Fredman, "The positive right to security," in *Security and human rights*, B. J. Goold, L. Lazarus, Eds. Oxford: Hart Publishing, 2007.

[10] A. Walsham, "The social history of the archive: Record-keeping in early modern Europe," *Past & Present*, vol. 230, no. suppl_11, pp. 9–48, 2016, doi: 10.1093/pastj/gtw033.

[11] M. Brosius, "Ancient archives and concepts of record-keeping: An introduction," in *Ancient archives and archival traditions: Concepts of record-keeping in the ancient world*, M. Brosius, Ed. New York: Oxford University Press, 2003.

[12] S. Simitis, "Einleitung," in *Kommentar zum Bundesdatenschutzgesetz (BDSG)*, S. Simitis, U. Dammann, O. Mallmann, H.-J. Reh, Eds. Baden-Baden: Nomos Verl.-Ges., 1978.

[13] D. Markopoulou, V. Papakonstantinou, P. de Hert, "The new EU cybersecurity framework: The NIS Directive, ENISA's role and the General Data Protection Regulation," *Computer Law & Security Review*, vol. 35, no. 6, 2019, doi: 10.1016/j.clsr.2019.06.007.

[14] European Union. (2013). *Cybersecurity Strategy of the European Union: An open, safe and secure cyberspace*, JOIN/2013/01 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52013JC0001. [Accessed: Sep. 1, 2022].

[15] M. Baezner, S. Cordey. (2019, Mar. 3). *National cybersecurity strategies in comparison – Challenges for Switzerland*, Zürich: Center for Security Studies (CSS), ETH Zürich, doi: 10.3929/ethz-b-000352773.

[16] S. Dimitrova, S. Stoykov, Y. Kochev, "National cybersecurity strategies in Member States of the European Union," *ACJ*, vol. 4, no. 73, p. 54, 2015, doi: 10.17770/acj.v4i73.4355.

[17] European Union. (2020). *The EU's Cybersecurity Strategy for the Digital Decade, European Commission*, JOIN(2020) 18 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=JOIN:2020:18:FIN. [Accessed: Sep. 1, 2022].

[18] European Network and Information Security Agency. (2012, Dec. 19). *National cyber security strategies.* [Online]. Available: https://www.enisa.europa.eu/publications/national-cyber-security-strategies-an-implementationguide. [Accessed: Sep. 1, 2022].

[19] European Network and Information Security Agency. (2016). *NCSS good practice guide: Designing and implementing national cyber security strategies.* [Online]. Available: https://data.europa.eu/doi/10.2824/48036. [Accessed: Sep. 22, 2022].

[20] European Network and Information Security Agency. (2020). *National capabilities assessment framework.* [Online]. Available: https://data.europa.eu/doi/10.2824/590072. [Accessed: Sep. 22, 2022].

[21] European Union Agency for Cybersecurity (2019). "Good practices in innovation on cybersecurity under the NCSS" [Online]. Available: https://data.europa.eu/doi/10.2824/01007. [Accessed: Sep. 22, 2022].

[22] A. Jacuch, "Comparative analysis of cybersecurity strategies", On-line Journal Modelling the New Europe, no. 37, p. 102, 2021, doi: 10.24193/OJMNE.2021.37.06.

[23] E. Luiijf, K. Besseling, P. de Graaf, "Nineteen national cyber security strategies," *International Journal of Critical Infrastructures*, vol. 9, no. 1–2, pp. 3–31, 2013, doi: 10.1504/IJCIS.2013.051608.

[24] G. Christou, *Cybersecurity in the European Union: Resilience and adaptability in governance policy*. Basingstoke, New York: Palgrave Macmillan, 2016.

[25] M. Dunn Cavelty. (2013). *A Resilient Europe for an open, safe and secure cyberspace, UI Occasional papers.* [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2368223. [Accessed: Sep. 9, 2022].

[26] C. Moorey, *A History of Crete*. London: Haus Publishing, 2019.

[27] N. Adiyeke, N. Adiyeke, E. Balta, "The Poll Tax in the years of the Cretan War: Symbol of submission and mechanisms of avoidance," *Thesaurismata*, vol. 31, pp. 323–59, 2001.

[28] V. Papakonstantinou, "Cybersecurity as praxis and as a state: The EU law path towards acknowledgement of a new right to cybersecurity?," *Computer Law & Security Review*, vol. 44, 2022, doi: 10.1016/j.clsr.2022.105653.

Digital Sovereignty Strategies for Every Nation

Ali Shoker | **Resilient Computing and Cybersecurity Center (RC3), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST)**, ORCID: 0000-0002-4898-9394

Abstract

Digital Sovereignty must be on the agenda of every modern nation. Digital technology is becoming part of our life details, from the vital essentials, like food and water management, to transcendence in the Metaverse and Space. Protecting these *digital assets* will, therefore, be inevitable for a modern country to live, excel and lead. Digital Sovereignty is a strategic necessity to protect these *digital assets* from the monopoly of friendly rational states, and the threats of unfriendly Malicious states and behaviors. In this work, we revisit the definition and scope of digital sovereignty through extending it to cover the entire value chain of using, owning, and producing *digital assets*. We emphasize the importance of protecting the operational resources, both raw materials and human expertise, in addition to research and innovation necessary to achieve sustainable sovereignty. We also show that digital sovereignty by autonomy is often impossible, and by mutual cooperation is not always sustainable. To this end, we propose implementing digital sovereignty using Nash Equilibrium, often studied in Game Theory, to govern the relation with Rational states. Finally, we propose a digital sovereignty agenda for different country's digital profiles, based on their status quo, priorities, and capabilities. We survey state-of-the-art digital technology that is useful to make the current *digital assets* sovereign. Additionally, we propose a roadmap that aims to develop a sovereign digital nation, as close as possible to autonomy. Finally, we draw attention to the need of more research to better understand and implement digital sovereignty from different perspectives: technological, economic, and geopolitical.

Keywords

autonomy, digital sovereignty, digital strategy, Nash Equilibrium, sovereign technology

Corresponding author:
Ali Shoker, King Abdullah University of Science and Technology (KAUST), 23955–6900 Thuwal, Kingdom of Saudi Arabia;
ORCID: 0000-0002-4898-9394;
ali.shoker@kaust.edu.sa

Cite this article as: A.Shoker, "Digital Sovereignty Strategies for Every Nation," ACIG, vol. 1, no. 1, pp. 69–88, 2022. DOI: 10.5604/01.3001.0016.0943

1. Introduction

Digital sovereignty is becoming a compelling priority to "control the present and the destiny" of modern nations [1], and a "make-or-break" issue [2]. It is a necessity for state's independence and national security in face of the increasing threats on *digital assets*. These threats are shown to be caused by both *like-minded* countries (e.g., USA, Germany, Canada, and Brazil) [3–7] and *non-like-minded countries* (e.g., USA, Russia, China, and Iran) [2, 8–10]. Protecting the *digital assets* from these threats, whether rational or malicious[1], is vital for modern countries given the unprecedented invasion of digital technology to our daily life essentials. The lack of digital sovereignty can undermine the automated and smart digital systems like water systems, food supply, smart power grids, telecommunications, Internet of Things, economy, health, governance, security, and defense [10–14]. Therefore, if these *digital assets* are threatened and compromised, national security will be at stake and human lives will be endangered.

1 —— The primary goal of a Malicious adversary is to cause harm to the target; whereas the primary goal of a Rational adversary is to increase its own payoff (utility), regardless of the target. A Rational player is selfish, but not necessarily Malicious. See details in Section 3.

The recent geopolitical tensions show that digital sovereignty is an urgent aspect more than ever. For instance, the recent bans of the USA on Huawei were mainly to protect the former's national digital sovereignty from undermining the telecommunication infrastructure and data [15, 16]. The tensions between the USA and China on the Taiwanese front are caused by the desire to control (around 70%) of the deep-tech semiconductor market fabrication [17]. The *Cyberwar* of state-backed malicious groups on websites in Finland, Italy, Romania, Germany, Norway, Lithuania, Czechia, Latvia, and elsewhere, is unprecedented [12]. Misinformation and systematic infiltration of social media can influence elections and democratic processes [2, 18]. Last, but not least, friendly relationships between countries cannot prohibit the surveillance of the officials of leading countries, like Germany [3].

Despite its importance, the study of Digital Sovereignty is still in its infancy. Since it touches upon different research areas like geopolitics, technology, and economics, more rigorous research efforts on the topic are still needed to fully understand the topic in a comprehensive and exhaustive way [18–20]. Digital sovereignty was originally promoted as data privacy and ownership – driven by political propagandas for "internal legitimacy" – to please the citizens [3, 18, 19, 21]. Then, it was defined and studied in different technological sovereignty areas, including digital, network, data, spectrum, Internet, cybersecurity, computer, and information [3, 22, 23]. Recently, the definition of digital sovereignty has got more attention in two dimensions. The first is on economic monopoly and intellectual property, that mainly targets the semiconductor fabrication, 5G infrastructure, and misuse of Artificial Intelligence (AI) with data [2, 18, 24–27]. The second dimension is related to the Cyberwar [14, 19, 28–30]. Although these works emphasize important perspectives, they lack a definition that captures the big picture and, thus, leading to an incomplete perception of the issue, while leaving gaps in approaching it.

We therefore introduce a comprehensive definition of digital sovereignty that covers the entire value chain of a state's *digital assets*. Our definition (discussed in Section 2) captures the entire scope of the digital sovereignty's spanning data, infrastructure, fabrication, raw material,

operational resources (raw material and humans), and research & innovation. The latter three aspects are noteworthy because they are often discarded or underestimated in literature. Several countries outsource their data and security operations (i.e., giving up digital sovereignty) to a handful worldwide known companies because of the lack of trained operational expertise and subject matter knowledge [1, 19, 31, 32]. China and Taiwan are leading the 5G and semiconductor manufacturing sectors because of their research advancements [15–17]. Similarly, the EU is lagging behind the USA and China in AI due to the lack of the highest *caliber* of talent and enough investment in R&I [32, 33]. On the other hand, reality shows that the shortage in energy supply as operational material can lead to a major digital shutdown, as in Ukraine and Lebanon [34–36]. We discuss the scope of digital sovereignty in detail, accompanied with a threat analysis to highlight the relevance and severity on these aspects.

The means to achieve digital sovereignty is even more challenging. We highlight the two main strategies proposed in literature and practice: sovereignty by *autonomy and by cooperation*. We show that although autonomy is the most effective sovereign way that states should seek, it is *impossible* with the current geopolitical landscape. This gap is often bridged with the cooperative treaties and alliances, e.g., following the cooperative bargaining problem studied in Decision Theory and *Game Theory* [37, 38]. Nevertheless, we notice that the digital sovereignty problem incurs a notion of threat par excellence, which we address by suggesting another strategy called *Nash Equilibrium* Sovereignty.

Nash Equilibrium Sovereignty follows the *Nash Equilibrium* strategy in *Game Theory* [7, 38], which is a hybrid cooperative and non-cooperative strategy, in contrast to the cooperative *bargaining* strategy. *Nash Equilibrium* allows two players to converge to a stable situation without direct intentions to cooperate. Using *Nash Equilibrium* is reasonable since it is tailored, by definition [38], to games where some notion of threat is present, similar to the digital sovereignty *game* in our study. In particular, it targets the case of rational players, which well represents the modern state governance. This strategy should precede bargaining whenever possible, since it is more sustainable and guaranteed given the unexpected tensions that can arise between friend states or allies [3, 16, 40, 41]. For instance, the best strategy for two neighboring (although not-like-minded) states is to cooperate on Internet packet delivery. The best strategy for a *deep-tech* producer and a corresponding rare material producer is to exchange their production.

We argue that digital sovereignty should be on the agenda of every modern government that embraces the digital world. While this topic is indeed among the top priorities of some states, like the EU and USA [14, 32, 42–44], many non-developed or developing countries consider themselves unconcerned, either because of underestimating its impact, or considering it a dream – due to the lack of capacities to implement it. We alleviate these misconceptions by introducing a digital *sovereignty agenda for every nation*, considering three country profiles that represent the digital status quo of all countries: *User*, for countries that manly outsource data; *Owner*, for countries that purchase and own infrastructure; and *Producer*, for countries that manufacture or develop digital technology. The producer level is the

ultimate target of sovereign states because it leads to digital autonomy and reduces external dependencies. The proposed agenda stands as a high-level roadmap for governments to (1) ensure an attainable level of digital sovereignty defined according to its posture, and (2) lift its profile to the most ambitious level, i.e., the Producer. As a case study, we drive a non-exhaustive survey of the recent technological techniques and security countermeasures that can be used in implementing digital sovereignty. We then demonstrate which means, among these techniques, can be used by the three profiles to protect their prioritized *digital assets*.

Our conclusion draws attention to the lack of enough studies on understanding digital sovereignty, especially those that study the interplay between technology, economy, and geopolitics. We particularly encourage further research on the *means* used to achieve it.

2. A Comprehensive Definition and Scope of DigitalSovereignty

The recent interest in digital sovereignty led to many definitions in the three worlds: public, technology, and geopolitics [2, 3, 18–20, 24–27]. Unfortunately, none of those are sufficiently comprehensive due to their focal perspective, which may impede the comprehensive and exhaustive study & implementation of digital sovereignty. For instance, there has been a huge emphasis on data sovereignty, that often restricts digital sovereignty to data privacy and ownership, targeting citizen's legitimacy, i.e., pleasing the people [3, 18, 19, 45]. Many technological sovereignty definitions have considered the sovereignty of different technological fields like Technological sovereignty, Digital sovereignty, Network sovereignty, Data sovereignty, Spectrum sovereignty, Internet sovereignty, Cyber sovereignty, Computer sovereignty, Network sovereignty, and Information sovereignty [3, 22]. Geopolitical definitions got inspired by *State Sovereignty* [3, 46] and focused on selected digital facets in the realm of the Cyberwar or the monopoly of resources [14–16, 19, 28–30]. We bridge this gap by providing a comprehensive definition to digital sovereignty as follows:

2.1. Definition 1

Digital Sovereignty of a state is possessing the supreme authority over all its *digital assets*, including the entire value chain: data, infrastructure, operations, supply chain, and knowledge.

The above definition is derived from the definition of *State Sovereignty* [3, 46] and applied to the state's "*digital assets*". The salient novelty in our definition is considering the entire digital value chain of a digital asset. Although the most valuable digital asset is the "data" itself; it could be undermined if digital sovereignty does not address the entire value chain that generates, stores, processes, operates, and manages the data – as we explain next.

State Sovereignty as an inspiration. State Sovereignty is commonly defined as possessing the supreme authority over a territory [3, 46]. This

authority is manifested as having full control over the territory. This definition should not, however, restrict the term "territory" to the landmass of a state; it rather includes the entire assets lying above and underneath, like the people, animals, space, air gases, oils, minerals, etc [47, 48]. In this sense, it is understood that the "territory" here represents the entire collection of assets that a state possesses (even if they reside abroad).

Data, the core digital asset. We argue that the entirety of the existence and importance *digital assets* are for the sake of "data", deemed here as the "core digital asset". Data is the digital[2] representation form of any piece of information. Data can have an immense impact on the state that grows as the reliance on digital technology grows; it is, therefore, inevitable for any modern and developing country. Data can be processed and presented as useful insights to make thoughtful and actionable decisions, or used to autonomously control other vital Cyber-Physical assets like power grids, water distribution, transportation, smart factories, etc. Unfortunately, experience shows that compromising these systems can endanger national security, human lives, and democracy [2, 11, 18].

2 —— Digital data is, typically, binary values of a physical quantity such as voltage or magnetic polarization.

The "borders", or lack thereof. Nevertheless, the "borders" in our definition to digital sovereignty are softer than those defined in State Sovereignty. As Barlow explained more than two decades ago, the "[c]yberspace does not lie within your borders" [49]. In fact, modern countries make an extensive use of the Internet and other digital communications within and outside their physical borders, e.g., for social, economic, military, and governance matters. Without these global communication channels, it is not difficult to figure out how slow and constrained the governance, life, and economy would be in this era. Nevertheless, data that is transported off-borders often uses other's communication channels, stored in remote storage, and processed using remote processors and software. These are all non-state-owned digital technology over which the data-owning state has little to no control. This is analogous – though more complex – to controlling other state assets oversees, e.g., ships and diplomatic representations in other countries. Likewise, data within borders can also be compromised by physical intruders, spies, thefts, and cyberattacks [8, 9, 13, 50].

2.2. Scope: the value chain, beyond data

Data is futile if not stored, processed, and transported. This is only possible through maintaining a large and complex value chain that is key in defining the correct scope of the digital sovereignty. We sketch this scope visually in Fig.1. The scope includes seven domains or aspects that cover the entire digital value chain. Data, infrastructure, fabrication, raw material, operational material, research & innovation (R&I), and operations. The latter two domains span the entire spectrum of the former five stacked domains. As explain earlier, the main emphasis of literary definitions was on the software and hardware stacks underlying the data, with partial focus on fabrication and material. Nevertheless, the operational material, R&I, and operations, have got little notice despite their essential role, compulsory to *digital assets*.

We demonstrate the importance of the seven domains in Fig.1, highlighting the corresponding threat model for each. The threat model considers both the case of *rational and malicious threats. The former mainly addresses the economic monopoly of digital technology and intellectual property. The latter is studied in the light of the well-known CIA triad: Confidentiality, Integrity, and Availability* [51, 52]. In a nutshell, confidentiality specifies that unauthorised users cannot access or disclose the data, mainly due to privacy or intellectual property reasons. Integrity ensures that data at-rest or in-transit is genuine and consistent, i.e., not tampered with by unauthorized users. Availability ensures that data is always available to be read and/or updated by authorised users. There are several ways to violate these properties; some of which we discuss in the following points.

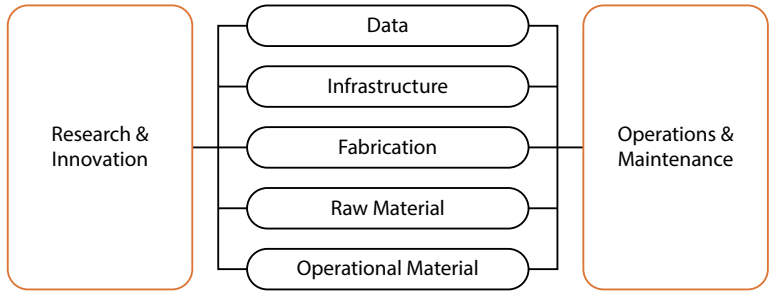


Figure 1. The architecture of the modern Internet, adapted from [4].

- **Data.** *Who rules the data, rules the world*. Data is the most valuable digital asset for which the entire value chain exists. It can hold sensitive state secrets, critical cyber-physical operational data, (bio)medical data, social data, and meta-data (i.e., data of data)[3]. Data is prone to threats on the three CIA triad [51]. Nevertheless, confidentiality can be easily compromised if it is outsourced for storage or computation outside the state's territory. Although a data operator may protect the data from unauthorized access, the data is under its mercy [1, 3]. This adversary (operator) is often called trusted-but-curious. This is why many countries have set regulations to protect outsourced data. Availability is mainly violated by *Denial of Service (DoS)* attacks or even encrypted for ransom reasons. For instance, the *WannaCry* ransomware attack has more than 230,000 devices [8, 50]. Unfortunately, many of these cyberattacks are believed to be coordinated by armed or hired groups by governments [8, 53].

- **Infrastructure (digital).** This is the set of software and hardware without which data cannot be: maintained, stored, processed, or transported. Digital infrastructure plays a main role in transforming data into useful information. It can be software such as: operating systems, protocols, AI algorithms, routing protocols, drivers, office tools, search engines, websites, etc. It can also be hardware that represents computers, servers, routers, cables, embedded devices, Integrated Circuits (ICs), etc. The application scope is also immense, spanning social networks, mobile applications, browsers, operating systems, propriety design, diagnostic tools, simulators, and testbeds. The threats on software and hardware can violate digital sovereignty

3 —— For details see New Urban Communities Authority (newcities.gov.eg/english/New_Communities/badr/default.aspx).

in many ways. A malware or a backdoor in an imported software can compromise the three CIA properties [51]. In the last decade, we noticed more resistance to stop the hegemony of US-based software. There are continuous calls and attempts in the EU to find rivals to the USA's GAFAM (Google, Amazon, Facebook, Apple, Microsoft) and Chinese BAT (Baidu, Alibaba, and Tencent) companies [1, 26, 32, 54]. Until recently, the USA is also suffering from serious infrastructural threats. For instance, in one year, 649 critical infrastructure entities had been attacked according to the FBI [11]. *SolarWinds Orion Platform* has also been subject to one of the largest infrastructure's backdoor attacks in the USA [10]. These attacks can even be more critical, like the famous *Stuxnet* malware on the Iranian nuclear-fuel plant [13]. Finally, although the attack on the Russian *Yandex Taxi* remained mild and contained, it represents a simple example of what could happen if the semi-autonomous transportation infrastructure is attacked [12].

- **Fabrication.** This is the most complex part in the value chain as it needs deep tech and rare raw materials. Hardware is rarely manufactured by one country and can be computers, mobiles, telecommunication equipment, datacentres, vehicles, medical equipment, smart devices, manufacturing control devices, modern vehicles, etc. In particular, silicon-hardware fabrication, e.g., semiconductors, requires decent technological advancements and industrial capabilities that exist in a handful of countries [16, 17]. This is critical from a digital sovereignty perspective since a producing country can ban any item in the supply chain, causing serious impacts on other countries [15, 16]. This is an economic and political threat of paramount significance. For instance, the sanctions on Russia are thought to cause a shortage in semiconductor chips that forced the Russian military to reuse chips from dishwashers and refrigerators [25]. *Huawei* lost its position at the top of mobile manufacturers due to bans by USA, Australia, New Zealand, Japan, and Taiwan [15, 16]. We also shed the light on the most serious attacks from a sovereignty perspective: trojans and backdoors [55, 56]. These are critical by being not easily detectable, and the producer is often supposed to be trusted. A state can buy a compromised hardware, e.g., military device, with a backdoor trojan that can be controlled by the producer/vendor [56]. A trojan can also play the "kill Switch" and, thus, activate itself after some time to ambiguate the attack for the buyer. The Israeli attack on the Syrian radar in 2007 is thought to be caused by a kill switch in an imported radar [56]. Similarly, it is extremely hard to verify if kill switches exist in other exported military weapons, like aircrafts. Nevertheless, these backdoors do exist in reality, like the one found in an Alcatel's military-grade field-programmable gate array (FPGA), although the company denies prior knowledge of it [57].

- **Raw material.** Fabrication requires a huge number of essential raw material, without which the product cannot be realized. The most typical example is semiconductor's resources that includes precious solids like *Gold, Aluminium, Diamond, Gallium Nitride and Silicon Carbide* [58–61]. These materials are used to produce the main hardware fabric primitives *Diodes, Transistors, Thyristors*, and other ICs [60, 62–64], using more than one hundred scarce and Noble gases, produced in very few countries [60, 65,

66]. The economy of these materials is subject to perturbations or shortage due to tensions between countries. For instance, the USA estimates that a Chinese invasion of Taiwan could disrupt the world economy and yield a loss of more than a trillion US dollars [17]. The recent Russian-Ukrainian war is not far from this monopoly, as Ukraine is a major producer of many noble gases – critical for semiconductor fabrication. The authors [67, 68] claim that if Russia retains control of *Mariupol* (a major Ukrainian factory where many Nobel gases like Neon, Krypton, Xenon, and Helium are produced) and restarts the city's damaged plants, 95% of the market could wind up in the hands of Russia and China.

- **Operational material.** These are raw materials needed for the operation of the infrastructure like energy and cooling sources, e.g., gas, oil, hydrogen, sunlight, wind, air, ice, etc. The entire value chain becomes useless if these operational resources are not guaranteed. For instance, the recent power outages due to the war in Ukraine and the Lebanese economic crisis led to major cut offs in the Internet and telecommunication industries [34–36]. On the other hand, Nordic countries are attracting dozens of major world-wide data centres due to the natural ice cooling [69]. This makes it more appealing and affordable to export data to these datacentres, but can also undermine data sovereignty as we explained above.

- **Research and Innovation (R&I).** This is a strategic aspect since human and institutional resources guarantee sustainable digital sovereignty through the quest for novel solutions and innovations [32, 70]. Reality shows that even leading countries can lose their position as a consequence of the lack of advancement in R&I. For instance, Taiwan's leadership in semiconductor fabrication is a result of consistent research and investment [32]. The Chinese leap in 5G and AI is referred to the higher percentages of R&D employees, as reported by Goldman Sachs [32]. The lag of the EU in AI innovation gives the impression that "China and the US innovate, while Europe can only regulate" [70].

- **Operations and Maintenance.** This includes the trained human resources with the subject-matter expertise to be able to operate the infrastructure. The lack of these trained resources can lead to sovereignty issues due to the need to outsource the data, or hiring third-party entities to maintain the operations [31]. Outsourcing sensitive data or cybersecurity data can compromise the countries' digital sovereignty since confidentiality and integrity would be violated as discussed earlier [26, 31]. Unfortunately, the worldwide regulations GDPR, CLOUD, among others [18, 21, 71, 72], fall short to mitigate these threats fully.

This wide scope and threat model indicates the complex nature of ensuring digital sovereignty at all levels. This represents a challenge in finding the correct means to ensure digital sovereignty. In addition, the above study shows that these means are not only technical, but rather economic, educational, and geopolitical by nature. This requires following economic strategies, geopolitical diplomacy and using special techniques and countermeasures to ensure data sovereignty, as we show next.

3. Strategies for Digital Sovereignty

The means to address digital sovereignty is cumbersome given its wide scope and the threats discussed above. The recent surveillance, leaks, threats, and incidents [3–6, 8, 9, 18, 50] called for compelling research to study the approaches towards sovereignty [16, 18, 22, 32, 43, 44, 46, 54, 73]. Two main directions are being heavily investigated: (1) autonomy that tries to build on self-reliance and independence to reduce the extremal influences [16, 18, 43], and (2) partnership that builds on cooperative bargaining through allies and coalitions to bridge the individual state gaps [16, 18, 28]. We show here that although both methods are useful, they fall short to achieve the sought digital sovereignty completely. Therefore, we propose using a complementary way that is more effective and reliable for *rational* behaviours, like countries. This approach is a hybrid cooperative and on-cooperative Game Theoretic model using *Nash Equilibrium*, introduced for rational behaviours by design [39, 49, 74].

3.1. Sovereignty by Autonomy

This approach promotes the independence and self-reliance means to defend against external threats [2, 16, 18, 43]. This should be the primary strategy used at the entire scope. While some purely technical domains like data, infrastructure, and operations are possible using state-of-the-art technology (as discussed later), other domains like fabrication and raw material seem more challenging. Although theoretically sound, this approach is often impractical for these domains with the current geopolitical landscape. Our conjecture is that digital autonomy is especially impossible in the case of small states. The reason is that the wide scope of digital sovereignty makes it very unlikely to autonomously supply and maintain the entire supply chain. Practice shows that even the digital sovereignty of leading and large countries, like the USA and China, can be undermined by the dependency on smaller or developing countries, like Taiwan [17]. While these large states can cope with it, this approach is undesired since it conflicts with other interesting properties like resilience [14, 18, 75], that embraces vendor diversity, economic openness to global consumers, and leadership that explores the best research ideas and minds worldwide [2, 24, 76]. It is noteworthy to mention that this approach may eventually lead to isolation and conservative relations with peer states, which can have a negative impact on the global and national technological advancements and national leadership. Note that the notion of "strategic autonomy" [18] is a little different from the autonomy we use here, as it may also refer to autonomy in decision making as well, which can itself follow several approaches, e.g., autonomy and mutual cooperation.

3.2. Sovereignty by Bargaining (i.e., cooperation)

This approach is complementary to the digital sovereignty by autonomy. It is based on building necessary partnerships with other states to bridge

a gap in the national digital value chain [16, 18, 28]. Bargaining takes a form of unilateral treaties between states, that leads to mutual benefits, or multilateralism form where coalitions and alliances are built for the common benefit of the group, e.g., the EU, NATO, or Gulf Cooperation Council (GCC).

Technically, this approach is realized by solving the cooperative *bargaining* problem, studied in *Decision Theory and Game Theory* [37, 38]. Classical bargaining is based on negotiations where a decision *utility* (i.e., representing payoffs) of expected values in the future is maximized, as suggested in the *Neumann–Morgenstern (VNM)* utility theorem in 1953 [77]. More formally, a two-person bargaining problem consists of a pair *(F, d)*, where *F* represents the set of feasible agreements and *d* is the disagreement point (payoffs) if bargain terminates without an agreement. The solution is to find a function *f* that takes a bargaining problem *(F, d)* as input and returns a feasible agreement as outcome, i.e., $f\ (F, d) \in F$. Finding the solution *f* can follow different criteria, like maximizing the product of gains (i.e., *Nash Product*) [37, 38], equalizing the gains [49], among others [67, 74].

Nevertheless, bargaining has been theoretically criticized for being unrealistic for several reasons [39, 47, 74]. The most relevant reasons in our context are (1) assuming a coalition of all states, e.g., relevant to a digital domain, will form; and (2) it ignores the effects of external actions (from other states) to the coalition. In the digital world, these assumptions are unrealistic given the wide dimension, and the evolving nature of digital technology that cannot be defined or restricted. For instance, despite the dominance of the USA in software, hardware, and telecommunications, the Chinese TikTok and Huawei's 5G have made a breakthrough in these domains [15, 16, 32, 33]. On the other hand, extending the coalition by joining new relevant states is not always successful as per the current geopolitical landscape. For instance, the attempts to strengthen cooperation with Taiwan is recently yielding heated dispute between the USA and China [17]. In the EU, the membership of states with digital manufacturing capabilities, like Turkey; or Nobel gases production, like Ukraine, are witnessing resistance [78]. The same holds for the NATO's membership of Sweden and Finland [79]. The *EU's Brexit* is another typical example – not only at the digital front though – showing that even an existing membership may break [16]. Finally, there is an increasing Chinese bilateral "third country" [16, 40] influence on central and *eastern Europe* (e.g., Italy, Hungary, Slovenia, and Greece). This undermined the EU's unity in the realm of the Chinese *16+1 initiative, Belt and Road, and China-CEEC* [40, 41].

3.3. Sovereignty by Nash Equilibrium

We propose using the *Nash Equilibrium* strategy studied in *Game Theory* as a hybrid cooperative and non-cooperative strategy to digital sovereignty when autonomy is not viable [38, 39, 49, 74]. Our inspiration is referred to the nature of the digital sovereignty problem that, by definition, incurs a notion of threat between *rational* players. State governance is rational by excellence and the *digital assets* are highly prone to several threats,

as described in the previous section. Our observation is that digital sovereignty by bargaining is not always realistic since it becomes a "solution applied to a wrong model", where states are assumed to be *benign* or even *altruistic*. This observation is consistent with the experimental studies on various bargaining models [47] showing that bargainers are found to focus on conceptually easy solutions that are beneficial to both parties. Nevertheless, bargaining is still a useful tool when sovereignty by *Nash Equilibrium* is infeasible.

Technically, a chosen strategy (of a set of actions) among possible ones is a *Nash Equilibrium* if no player can do better by unilaterally changing its strategy. More formally, let two players (i.e., states in our case) *A* and *B* have S_A and S_B as sets of possible strategies with utility functions (payoffs) u_A and u_B, respectively. A binary setting (s^i_A, s^j_B), where s^i_A s_A and sjB SB, is a *Nash Equilibrium* if A cannot obtain a higher utility payoff (u_A) than choosing s^i_A, i.e., $u_A(s^i_A, s^j_B) > u_A(s^i_A, s^x_B)$ for any $sxB \in SB$. The same holds by symmetry for player *B*, with respect to strategy *sjB* and utility *uB*. Since both A and B cannot do better, the game will stabilize, and the chosen strategies are enforced as if there is a cooperative agreement.

Practical examples on this approach are the cooperation of two *not-like-minded* neighboring states on Internet packet delivery. *Nash Equilibrium* is achieved since both states will deliver packets to their destination otherwise their own packets will be at stake. Another example in the semiconductor market is between the USA, that dominates the semiconductor design market [16], and Taiwan, that own the cutting edge 3nm and 5nm semiconductor fabrication [17]. Since available alternatives are scarce on both sides, none can efficiently produce semiconductors alone, which forces them to the *Nash Equilibrium* strategy. On the other hand, the wave of decentralized systems, inspired by Blockchains, are leading several use-cases in Fintech, Supply Chain, Cloud Computing, Governance, etc., that partially follow this rational model [80].

Sovereignty by *Nash Equilibrium* exhibits some drawbacks rooted in their design and application. Two drawbacks are particularly more relevant in our context. The first is the existence of multiple Nash Equilibria in one game, which sometimes prevents reaching the highest utility possible (i.e., called *Pareto optimality*) [39, 74]. The approach also assumes the knowledge of all states to all potential strategies, which might be too expensive or infeasible for small states. A *dummy* player who cannot attain this information may not be rationally on par with its counterpart, which violates the original assumption [39, 81].

Finally, we argue that the three strategies should be used together to ensure digital sovereignty. The preferred one must always be autonomy, followed by *Nash Equilibrium*. We recommend the latter over bargaining since it is more sustainable as discussed above. This is due to the intrinsic needs of the counterparts and, thus, imposing the *Nash Equilibrium* strategy can be seen as a *soft enforcement*. Bargaining can be an alternative to *Nash Equilibrium* when the latter is infeasible. Nevertheless, the cooperation between states is always encouraged in general, although it is ineffective when a rational player is not *playing fair*.

4. An Agenda for Digital Sovereignty

The above challenges indicate an unprecedented need for a digital sovereignty strategy for every nation. Nevertheless, it seems this represents a major concern for a limited number of countries [5, 15, 28, 32, 43]. The reason could be referred to underestimating the criticality of digital sovereignty or the lack of awareness. We have discussed earlier the criticality for the topic to every country embracing the digital world. In addition, non-developed countries may consider themselves unconcerned or not ready because of their limited capability to do anything about digital sovereignty.

To alleviate these misconceptions, we propose an agenda for the digital sovereignty targeting three profiles of nations, based on their capabilities and digital maturity. These profiles represent the majority of worldwide countries. The agenda includes an implementation of a state's current digital sovereignty as well as a future development plan. Being strategic, digital sovereignty follows a long duration roadmap that makes it an urgent priority for any modern government, sooner not later. We divide the agenda into three parts: self-assessment, planning, and implementation.

4.1. Self-assessment

1. *Establish a National Digital Sovereignty Agency (NDSA)*: this agency is in charge of the assessment, planning, implementation, and evaluation plans for digital sovereignty. It mediates the discussions with all other relevant agencies and ministries to *digital assets*. In particular, it can work under the supervision of a national state sovereignty agency if exists, and coordinates with other sovereignty agencies, e.g., food and borders sovereignty agencies.

2. *Identify national digital assets*: NDSA appoints relevant teams, workshops, and discussions to identify the state's *digital assets*.

3. *Drive a sovereignty threat and risk assessment:* all identified *digital assets* should be subject to a threat and risk assessment, covering the entire value chain discussed in the previous section.

4. *Designate the states' posture*: User, Owner, Producer. This gives the state a digital profile based on its status quo and capabilities. A posture of each digital asset can be assigned (more details below):

- *User:* mostly uses digital technology and infrastructure that others produce or own.

- *Owner*: mostly owns digital technology and infrastructure that others produce.

- *Producer*: mostly produce all used technology and infrastructure.

4.2. Planning

A plan for *short-, mid-, and long-term stages should be defined*. Digital sovereignty is a long-road project that clearly requires a long-term roadmap (e.g., 30 years, for a developing country). However, it should be developed incrementally with *best effort* over a *mid-term plan* (e.g., 10 years), and a *short-term plan* (e.g., 3 years). The mindset and goal are to *advance the posture from User to Owner or Producer*. We do not envision technical restrictions that prohibit the planning from User to Producer postures in some aspects, although it may be infeasible in others. The proposed workflow for the planning is as follows:

1. *Set goals based on the posture expected at each stage:* different *digital assets* may have different maturity levels (User, Owner, or Producer). This requires setting a national digital sovereignty goal for all *digital assets* at every stage. The purpose is to lift the digital maturity to a standard norm, preparing for the next stage.

2. *Set sector priorities based on its severity level*: it might not be feasible to focus on all sectors at once. A state may start with the most critical or sensitive sectors or assets.

3. *Adopt a sovereignty strategy*: the ultimate goal for nations is to achieve autonomy in all *digital assets* and sovereignty domains as in Fig. 1. This may not be achievable in some economies depending on digital maturity and geopolitical reasons. However, the goals should be set high enough for each state. In parallel, *digital assets* that rely on external Producers or Owners may follow a *Nash Equilibrium* strategy at first. As discussed in the previous section, this is because it is a more sustainable and reliable strategy for rational states. The rest of external dependencies can be sought through bargaining, e.g., building economic partnerships and alliances via diplomacy. Therefore, the three strategies should be used simultaneously.

4.3. Implementation

Define and enforce legislations and regulations: this is the very first actionable step that is needed to set the standards that national and international stakeholders should abide to. This should account for some time (e.g., few years) before enforcement, leaving a suitable window of time for stakeholders to prepare for a successful and smooth change.

Apply State-of-the-Art (SotA) techniques and countermeasures: this is the most technical part of the implementation. It makes use of SotA tools and techniques tailored to make *digital assets* sovereign, e.g., like Privacy, (Cyber)Security, and Resilience. These techniques should be researched and developed on a regular basis, otherwise some of them will be deprecated with time and fail against an evolving stronger adversary. For instance, security techniques that rely on classical cryptography, e.g., Public-Key Cryptography (PKC), may at some point be replaced with *post-Quantum* cryptography [82]. A high-level workflow would be as shown in Fig. 2.

4.4. Implementation case study

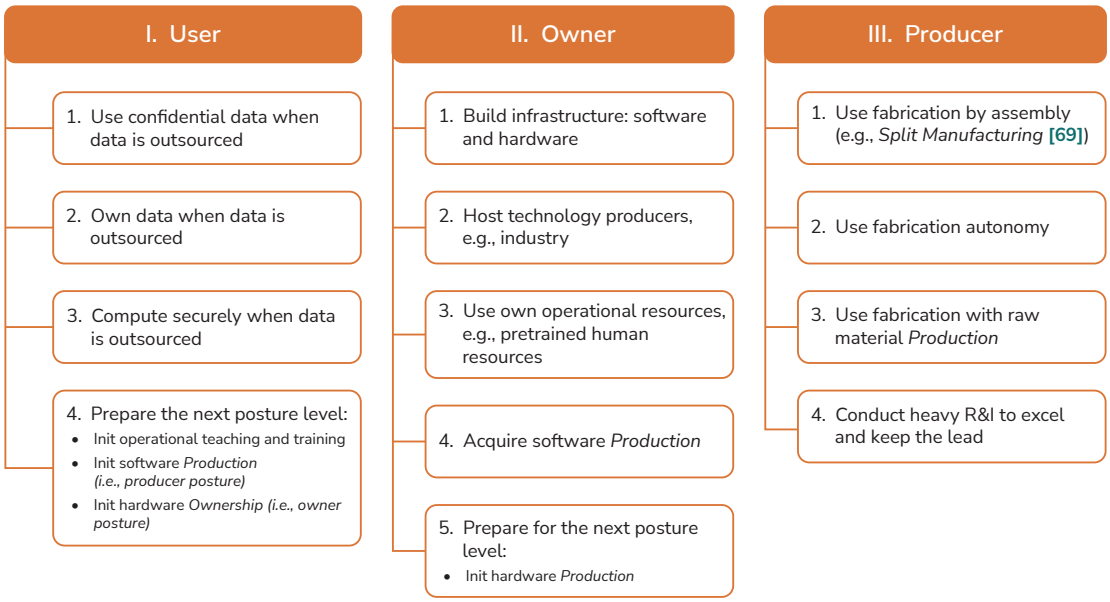


Figure 2. A high-level workflow for the implementation of digital sovereignty. The natural development plan is to acquire a User, then an Owner, then a Producer profile.

We present a case study to explain the methodology of the above implementation, showing that countries with different digital postures can do much on the sovereignty front. We demonstrate this in Fig. 3. and Tab. 1. The former conveys an example of selected state-of-the-art (SotA) techniques and countermeasures used for digital sovereignty. Although these techniques are part of SotA, the list is only meant to exemplify the state's capabilities and application feasibility – thus, it is not an exhaustive list. The right column represents five high-level techniques numbered from 1 to 5, mostly, but not strictly, in increasing level of complexity.

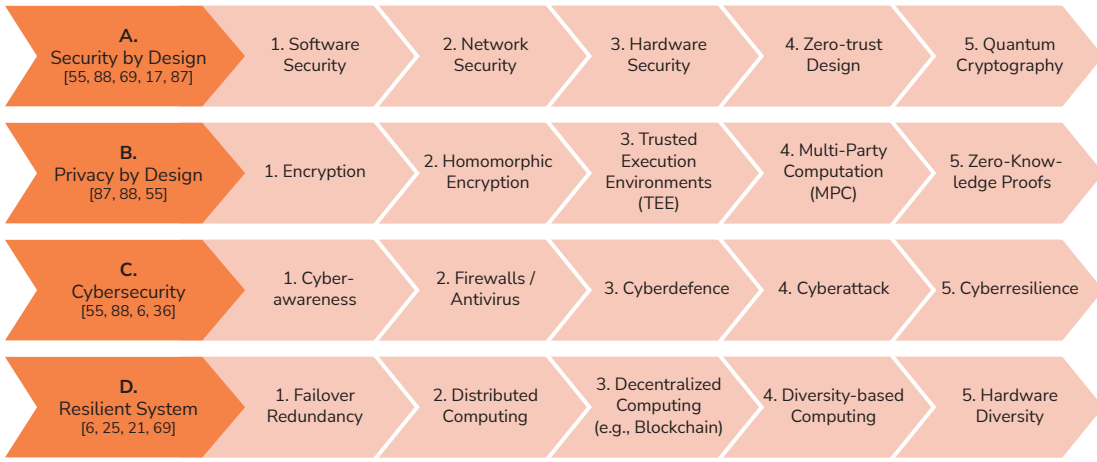


Figure 3. An example of state-of-the-art digital sovereignty techniques and countermeasures.

On the other hand, Tab. 1. matches the severity of applications or systems (in the left column) to the *state's posture* (in the bottom column). To be generic enough, we inspire the application severity from the "human needs" model in *Maslow's pyramid* [83, 84], with some exceptions for sensitive and deep technologies. However, a real-world case is to go over the entire digital-based services a government provide. Finally, we try to map the techniques used to the three defined postures. The main observation is that even states with (the lowest) User profile can improve their sovereignty with the currently available technology. More advanced postures, like the Owner and Producer have more capabilities to apply more techniques. We leave the details of this matching as an exercise to the reader, and we encourage a deeper study on this front.

Table 1. An example of matching the sovereignty techniques and countermeasures in Table 1 to a state's posture and severity levels (inspired by Maslow's pyramid of human needs [83, 84]).

Happiness	Entertainment	A.1–2	A.1–2	A.1–4
	Gaming	B.1	B.1	B.1–3
	Art	C.1–2	C.1–3	C.1–3
	Lesure		D.1–2	D.1–3
Needed	Enivornment	A.1–2	A.1–4	A.1–5
	Politics	B.1–2	B.1–3	B.1–5
	Education	C.1–2	C.1–3	C.1–5
	Social nets Associations		D.1–3	D.1–4
Critical	Air, Water, & Energy	A.1–2	A.1–5	A.1–5
	Security, Top Secret	B.1–3,	B.1–5	B.1–5
	Deep tech. (e.g., Space & Nuclear)	C.1–2	C.1–5	C.1–5
			D.1–4	D.1–5
		User	**Owner**	**Producer**

5. Conclusions

Digital sovereignty is getting more traction due to the data surveillance, leaks, cyberattacks, and monopoly of digital resources. In this work, we have introduced a new definition to digital sovereignty showing that its scope must include the operational material, human resources, as well as research and innovation. Indeed, the lack of any of these aspects makes outsourcing data storage, computation, or operation a must, which undermines sovereignty. We have provided a threat model to emphasize the criticality of all the aspects of the digital scope, giving real-world examples. Then we proposed a new digital strategy by *Nash Equilibrium* to be used when autonomy is not feasible, while cooperative bargaining with other states should not be discarded. We also proposed an agenda for digital sovereignty to set the roadmap for a higher profile digital maturity. We show that any country can improve its sovereignty by following available techniques and countermeasures.

Our work can benefit from different directions in the future. First, a more exhaustive threat model and countermeasures are needed. Second, the bargaining and *Nash Equilibrium* strategies require deeper study to make the implementation of digital sovereignty easier. In particular, it is interesting to study how to play these games within coalitions (e.g., EU and GCC) and alliances (e.g., NATO) without breaking the member state sovereignties. Third, the agenda we propose is high level; a lower-level roadmap that stands as a detailed template for governments to follow would be very useful. Finally, the sovereign techniques and countermeasures we survey can benefit from a more comprehensive and thorough research in the future [85–88].

References

[1] P. Bellanger, *La souveraineté numérique*, Paris: Institut Diderot, 2016.

[2] V. d. Leyen. (2021, Sep. 15). *State of the Union Address by President von der Leyen*, European Commission. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_21_4701. [Accessed: Oct. 24, 2022].

[3] S. Couture, S. Toupin, "What Does the Concept of 'Sovereignty' Mean in Digital, Network and Technological Sovereignty?," *SSRN Journal*, vol. 21, no. 10, pp. 2305–2322, 2018, doi: 10.2139/ssrn.3107272.

[4] A. Holpuch. (2013, Sep. 20). *Brazil's controversial plan to extricate the internet from US control, The Guardian*. [Online]. Available: https://www.theguardian.com/world/2013/sep/20/brazil-dilma-rousseff-internet-us-control. [Accessed: Oct. 19, 2022].

[5] J. A. Obar, A. Clement, "Internet Surveillance and Boomerang Routing: A Call for Canadian Network Sovereignty," TEM 2013: Proceedings of the Technology & Emerging Media Track – Annual Conference of the Canadian Communication Association, 2013, doi: 10.2139/ssrn.2311792.

[6] M. M. Lee, "The International Politics of Incomplete Sovereignty: How Hostile Neighbors Weaken the State," *International Organization*, vol. 72, no. 2, pp. 283–315, 2018, doi: 10.1017/S0020818318000085.

[7] *Brazilian Leadership and the Global Internet.* (2014, Apr. 28). AULA Blog. [Online]. Available: https:// aulablog.net/2014/04/28/brazilian-leadership-and-the-global-internet. [Accessed: Oct. 19, 2022].

[8] B. Jain. (2022, July 10). *Russia, China, North Korea and Iran lead in supporting aggressive cyber attackers, says HolistiCyber CEO*, The Times of India. [Online]. Available: https://timesofindia.indiatimes.com/world/rest-of-world/russia-china-north-korea-and-iran-leads-in-supporting-aggressive-cyber-attackers-saysholisticyber-ceo/articleshow/92779362.cms. [Accessed: Oct. 25, 2022].

[9] Congress. (2013, Mar. 20). *Cyber Threats From China, Russia, and Iran: Protecting American Critical Infrastructure*, U.S. Government Publishing Office. [Online]. Available: https://www.govinfo.gov/content/pkg/CHRG-113hhrg82583/html/CHRG-113hhrg82583.htm. [Accessed: Oct. 25, 2022].

[10] P. Paganini. (2021). *SolarWinds hack: the mystery of one of the biggest cyberattacks ever, Cybernews.* [Online]. Available: https://cybernews.com/security/solarwinds-hack-the-mystery-of-one-of-the-biggest-cyberattacks-ever/. [Accessed: Oct. 25, 2022].

[11] Industrial Safety and Security Source. (2022, Mar. 22). *Internet Crime Losses Surge in 2021: Report, ISSSource.* [Online]. Available: https://www.isssource.com/internet-crime-losses-surge-in-2021-report. [Accessed: Nov. 1, 2022].

[12] V. Petkauskas. (2022, Sep. 2). *Yandex Taxi hack creates huge traffic jam in Moscow*, Cybernews. [Online]. Available: https://cybernews.com/cyber-war/hackers-created-an-enormous-traffic-jam-in-moscow. [Accessed: Oct. 25, 2022].

[13] D. Kushner, "The real story of stuxnet," *IEEE Spectrum*, vol. 50, no. 3, pp. 48–53, 2013, doi: 10.1109/MSPEC.2013.6471059.

[14] *Cybersecurity and Infrastructure Security Agency*. (2022). CISA Strategic Plan 2023–2025. [Online]. Available: https://www.cisa.gov/sites/default/files/publications/StrategicPlan_20220912-V2_508c.pdf. [Accessed: Oct. 25, 2022].

[15] Statista Infographics. (2022, Jan. 30). *Infographic: Which Countries Have Banned Huawei?* [Online]. Available: https://www.statista.com/chart/17528/countries-which-have-banned-huawei-products. [Accessed: Nov. 1, 2022].

[16] D. Fiott, *Strategy and interdependence*, Paris: EU Institute for Security Studies, 2021.

[17] J. Leonard, D. Wu, K. Manson. (2022, Oct. 7). *Taiwan Tensions Spark New Round of US War-Gaming on Risk to TSMC*. [Online]. Available: https://www.bloomberg.com/news/articles/2022-10-07/taiwan-tensionsspark-new-round-of-us-war-gaming-on-risk-to-tsmc. [Accessed: Nov. 1, 2022].

[18] L. Moerel, P. Timmers. (2022, Jan. 1). *Reflections on Digital Sovereignty, Rochester*. [Online]. Available: https://papers.ssrn.com/abstract=3772777. [Accessed: Nov. 2, 2022].

[19] A. Obendiek, "Take back control? Digital sovereignty and a vision for Europe," *Jacques Delors Centre*, Policy Paper, 2021.

[20] G. Falkner, S. Heidebrecht, A. Obendiek, T. Seidl, "Digital Sovereignty – Rhetoric and Reality," *Online Conference*, 2022.

[21] European Commission. (2022, Oct. 17). *Europrivacy: the first certification mechanism to ensure compliance with GDPR, Shaping Europe's digital future*. [Online]. Available: https://digital-strategy.ec.europa.eu/en/news/europrivacy-first-certification-mechanism-ensure-compliance-gdpr. [Accessed: Oct. 25, 2022].

[22] C. Anupam, H. Sun, "Sovereignty 2.0," *Georgetown Law Faculty Publications and Other Works*, 2404, 2022.

[23] T.-H. Hu, *A Prehistory of the Cloud, MIT Press*. [Online]. Available: https://mitpress.mit.edu/9780262529969/a-prehistory-of-the-cloud. [Accessed: Oct. 19, 2022].

[24] V. d. Leyen. (2022, Sep. 16). *State of the Union Address by President von der Leyen*, European Commission. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_20_1655. [Accessed: Oct. 24, 2022].

[25] V. d. Leyen. (2022, Sep. 16). *State of the Union Address by President von der Leyen, European Commission*. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/speech_22_5493. [Accessed: Oct. 26, 2022].

[26] T. Nitot, N. Cercy. (2016). *Numérique: reprendre le contrôle, Framabook*. [Online]. Available: https://archives.framabook.org/numerique-reprendre-le-controle. [Accessed: Nov. 2, 2022].

[27] European Commission. (2019, Apr. 08). *Ethics guidelines for trustworthy AI, Shaping Europe's digital future*. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. [Accessed: Oct. 26, 2022].

[28] W. Goździewicz. (2019, Nov. 11). *Sovereign Cyber Effects Provided Voluntarily by Allies (SCEPVA), Cyber Defense Magazine*. [Online]. Available: https://www.cyberdefensemagazine.com/sovereign-cyber. [Accessed: Oct. 25, 2022].

[29] *Council on Foreign Relations, Tracking State-Sponsored Cyberattacks Around the World, Council on Foreign Relations*. [Online]. Available: https://www.cfr.org/cyber-operations. [Accessed: Oct. 25, 2022].

[30] H. Kjell, "Cyber resilience in firms, organizations and societies," *Internet of Things*, vol. 11, no. 1, 2020, doi: 10.1016/j.iot.2020.100204.

[31] M. Pańkowska, "Outsourcing Impact on Security Issues," in *Evolution and Challenges in System*

Development, J. Zupančič, W. Wojtkowski, W. G. Wojtkowski, S. Wrycza, Eds. Boston, MA: Springer US, 1999, pp. 235–246. doi: 10.1007/978-1-4615-4851-5_22.

[32] E. M. Leonard, J. Shapiro, "Strategic Sovereignty: How Europe can Regain the Capacity to Act," *European Council on Foreign Relations*, 2019.

[33] M. Tambiama, "Digital sovereignty for Europe," *European Parliamentary Research Service (EPRS)*, 2020.

[34] S. Moss. (2022, Mar. 3). *Ukraine: Mariupol Internet goes dark as power is cut off and city comes under siege, Data Center Dynamics.* [Online]. Available: https://www.datacenterdynamics.com/en/news/mariupol-ukraine-internet-attack/. [Accessed: Nov. 1, 2022].

[35] N. Durgham. (2022, Sep. 1). *Massive hike in cell service fees cuts off Lebanon's poor from the world, Washington Post.* [Online]. Available: https://www.washingtonpost.com/world/2022/09/01/lebanon-telecom-rates-poverty-crisis/. [Accessed: Nov. 1, 2022].

[36] K. Szulecki, *Conceptualizing energy democracy, Environmental Politics*, vol. 27, no. 1, pp. 21–41, 2018, doi: 10.1080/09644016.2017.1387294.

[37] J. F. Nash, "The Bargaining Problem", Econometrica, vol. 18, no. 2, 1950, doi: 10.2307/1907266.

[38] J. Nash, "Two-Person Cooperative Games", *Econometrica*, vol. 21, no. 1, pp. 128–140, 1953, doi: 10.2307/1906951.

[39] E. Maskin, "Commentary: Nash Equilibrium and mechanism design," *Games and Economic Behavior*, vol. 71, no. 1, pp. 9–11, 2011, doi: 10.1016/j.geb.2008.12.008.

[40] M. L. Wolff Jean Pisani-Ferry, E. Ribakova, J. Shapiro, Guntram. (2019, June 25). *Redefining Europe's economic sovereignty – European Council on Foreign Relations, ECFR.* [Online]. Available: https://ecfr.eu/publication/redefining_europes_economic_sovereignty. [Accessed: Oct. 29, 2022].

[41] *Secretariat for Cooperation between China and Central and Eastern European Countries, Cooperation between China and Central and Eastern European Countries.* [Online]. Available: http://www.china-ceec.org/eng. [Accessed: Oct. 29, 2022].

[42] L. Floridi, "The Fight for Digital Sovereignty: What It Is, and Why It Matters, Especially for the EU," *Philosophy & Technology*, vol. 33, no. 3, pp. 369–378, 2020, doi: 10.1007/s13347-020-00423-6.

[43] S. Anghel, "Strategic sovereignty for Europe," *European Parliamentary Research Service (EPRS)*, 2020.

[44] A. Braud, G. Fromentoux, B. Radier, O. Le Grand, "The Road to European Digital Sovereignty with Gaia-X and IDSA," *IEEE Network*, vol. 35, no. 2, pp. 4–5, 2021, doi: 10.1109/MNET.2021.9387709.

[45] T. Kukutai, J. Taylor, Eds., "Indigenous Data Sovereignty: Toward an agenda," *ANU Press*, 2016. doi: 10.22459/CAEPR38.11.2016.

[46] D. Philpott. (2003). *Sovereignty, Stanford Encyclopedia of Philosophy Archive.* [Online]. Available: https://plato.stanford.edu/archives/sum2016/entries/sovereignty. [Accessed: Oct. 19, 2022].

[47] J. A. Schellenberg, "Solving the Bargaining Problem," *Social Thought and Research*, 1990, doi: 10.17161/STR.1808.5046.

[48] M. Clark. (2022, Aug. 27). *Satellite-to-phone companies are thrilled about SpaceX and T-Mobile, actually, The Verge.* [Online]. Available: https://www.theverge.com/2022/8/27/23324128/t-mobile-spacex-satellite-tophone-technology-ast-lynk-industry-reactions-apple. [Accessed: Oct. 4, 2022].

[49] E. Kalai, M. Smorodinsky, "Other Solutions to Nash's Bargaining Problem," *Econometrica*, vol. 43, no. 3, pp. 513–518, 1975, doi: 10.2307/1914280.

[50] M. Savita, M. Patil, "A brief study of wannacry threat: Ransomware attack 2017," *International Journal of Advanced Research in Computer Science*, 2017.

[51] A. Rashid, H. Chivers, G. Danezis, E. Lupu, A. Martin, "The Cyber Security Body of Knowledge," *CyBoK.org*, 2019.

[52] W. Stallings, L. Brown, *Computer security: principles and practice*, Upper Saddle River: Pearson, 2012.

[53] A. d. Liedekerke, A. Laudrain. (2022). *Russia's Cyber War: What's Next and What the European Union Should Do, Council on Foreign Relations*. [Online]. Available: https://www.cfr.org/blog/russias-cyber-warwhats-next-and-what-european-union-should-do. [Accessed: Oct. 25, 2022].

[54] B. H. Bratton, "The Stack: On Software and Sovereignty." *The MIT Press*, 2016. doi: 10.7551/mitpress/9780262029575.001.0001.

[55] T. D. Perez, S. Pagliarini, "A Survey on Split Manufacturing: Attacks, Defenses, and Challenges," *IEEE Access*, vol. 8, pp. 184013–184035, 2020, doi: 10.1109/ACCESS.2020.3029339.

[56] IEEE Spectrum. (2008, May 1). *The Hunt for the Kill Switch*. [Online]. Available: https://spectrum.ieee.org/the-hunt-for-the-kill-switch. [Accessed: Nov. 1, 2022].

[57] S. Skorobogatov, C. Woods, "Breakthrough Silicon Scanning Discovers Backdoor in Military Chip," in *Cryptographic Hardware and Embedded Systems – CHES* 2012, vol. 7428, E. Prouff, P. Schaumont, Eds. Berlin, *Heidelberg: Springer Berlin Heidelberg*, 2012, pp. 23–40. doi: 10.1007/978-3-642-33027-8_2.

[58] R. F. Davis, "Critical evaluation of the status of the areas for future research regarding the wide band gap semiconductors diamond, gallium nitride and silicon carbide," *Materials Science and Engineering: B*, vol. 1, no. 1, 1988, doi: 10.1016/0921-5107(88)90032-3.

[59] *Diamond Foundry, Diamond Semiconductor Technology*. [Online]. Available: https://diamond-foundry.com/pages/diamond-semiconductor-technology. [Accessed: Oct. 23, 2022].

[60] D. P. Stockman, "Creating a semiconductor and the gases that make it happen," *Linde Electronics*, 2018.

[61] J. Boyd. (2022, Mar. 17). *This Diamond Transistor is Still Raw, But Its Future Looks Bright*, IEEE Spectrum. [Online]. Available: https://spectrum.ieee.org/this-diamond-transistor-is-still-raw-but-its-future-looks-bright. [Accessed: Oct. 23, 2022].

[62] Shindengen Electric, *What is N-type and P-type semiconductor?, Shindengen Electric Mfg. Co., Ltd*. [Online]. Available: https://www.shindengen.com/products/semi/column/basic/semi/rectifying_action.html. [Accessed: Oct. 23, 2022].

[63] Shindengen Group, *Basis of Semiconductors, Semiconductor, Shindengen Electric Mfg. Co., Ltd.* [Online]. Available: https://www.shindengen.com/products/semi/column/basic/semi/about.html. [Accessed: Oct. 23, 2022].

[64] All About Circuits, *The Basics of Power Semiconductor Devices: Structures, Symbols, and Operations*. [Online]. Available: https://www.allaboutcircuits.com/technical-articles/a-review-on-power-semiconductordevices. [Accessed: Oct. 23, 2022].

[65] S. Foster. (2022, Sep. 14). *Gallium oxide fuelling new chapter chip war, Asia Times*. [Online]. Available: https://asiatimes.com/2022/09/gallium-oxide-fuelling-new-chapter-chip-war. [Accessed: Oct. 26, 2022].

[66] NIST. (2009). *Index of Semiconductor Process Gases*. [Online]. Available: https://www.nist.gov/pml/sensor-science/fluid-metrology/database-thermophysical-properties-gases-used-semiconductor-0. [Accessed: Oct. 23, 2022].

[67] I. Kaminska, *Analysis: Noble Gases Are Suffering From Putin's War in Ukraine*, Washington Post. [Online]. Available: https://www.washingtonpost.com/business/energy/noble-gases-are-suffering-from-putins-war-in-ukraine/2022/05/19/594109ec-d731-11ec-be17-286164974c54_story.html. [Accessed: Oct. 26, 2022].

[68] G. Athanasia, G. Arcuri. (2022). *Russia's Invasion of Ukraine Impacts Gas Markets Critical to Chip Production,* The Center for Strategic and International Studies (CSIS). [Online]. Available: https://www.csis.org/blogs/perspectives-innovation/russias-invasion-ukraine-impacts-gas-markets-critical-chip-production. [Accessed Oct. 26, 2022].

[69] J. D. Christensen, J. Therkelsen, I. Georgiev, H. Sand, "Data centre opportunities in the Nordics," *Copenhagen: Nordic Council of Ministers*, 2018. doi: 10.6027/TN2018-553.

[70] N. A. Smuha, "The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence," *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, 2019, doi: 10.9785/cri-2019-200402.

[71] P. O. of the European Union. (2016, Apr. 27). 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) (Text with EEA relevance), Publications Office of the European Union. [Online]. Available: http://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1. [Accessed Oct. 25, 2022].

[72] DLA Piper, *Global Data Protection Laws of the World, World Map, DLA Piper.* [Online]. Available: https://www.dlapiperdataprotection.com. [Accessed: Nov. 2, 2022].

[73] P. Timmers, "The Technological Construction of Sovereignty," in *Perspectives on Digital Humanism*, H. Werthner, E. Prem, E. A. Lee, C. Ghezzi, Eds. Springer International Publishing, 2022, pp. 213–218. doi: 10.1007/978-3-030-86144-5_28.

[74] R. Serrano, "Fifty Years of the Nash Program, 1953–2003," *SSRN Journal*, 2004, doi: 10.2139/ssrn.724233.

[75] European Commission. (2022, Oct. 18). *Critical Infrastructure Resilience: stronger rules*. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_6238. [Accessed: Oct. 25, 2022].

[76] G. A. Schmidt Eric. (2022). *The U.S. Needs a Million Talents Program to Retain Technology Leadership, Foreign Policy*. [Online]. Available: https://foreignpolicy.com/2022/07/16/immigration-us-technologycompanies-work-visas-china-talent-competition-universities. [Accessed: Nov. 2, 2022].

[77] J. von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior: 60th Anniversary Commemorative Edition*, Princeton and Oxford: Princeton University Press, 2007, doi: 10.1515/9781400829460.

[78] *European Commission, Additional tools Candidate Countries and Potential Candidates*. [Online]. Available: https://ec.europa.eu/environment/enlarg/candidates.htm. [Accessed: Nov. 2, 2022].

[79] Reuters. (2022, May 19). T*urkey to reject Sweden and Finland's bid to join NATO*, Reuters. [Online]. Available: https://www.reuters.com/world/turkey-has-told-allies-its-no-sweden-finlands-nato-bid-erdogan-2022-05-19. [Accessed: Nov. 2, 2022].

[80] A. Joseph, M. Brunnermeier, "Blockchain economics," *Working Paper*, no. 25407, 2018.

[81] X. Jin, M. Dan, N. Zhang, W. Yu, X. Fu et al., "*Game Theory* for Infrastructure Security: The Power of Intent-Based Adversary Models," in *Handbook on Securing Cyber-Physical Critical Infrastructure*, S. K. Das, K. Kant, N. Zhang, Eds. Boston: Morgan Kaufmann, 2012, pp. 31–53. doi: 10.1016/B978-0-12-415815-3.00002-9.

[82] L. Malina et al., "Post-Quantum Era Privacy Protection for Intelligent Infrastructures," I*EEE Access*, vol. 9, pp. 36038–36077, 2021, doi: 10.1109/ACCESS.2021.3062201.

[83] S. Mcleod, "Maslow's Hierarchy of Needs," SimplyPsychology, 2018.

[84] A. H. Maslow, "A theory of human motivation," *Psychological Review*, vol. 50, pp. 370–396, 1943, doi: 10.1037/h0054346.

[85] P. Uday, K. Marais, "Designing Resilient Systems-of-Systems: A Survey of Metrics, Methods, and Challenges," *Systems Engineering*, vol. 18, no. 5, pp. 491–510, 2015, doi: 10.1002/sys.21325.

[86] A. Humayed, J. Lin, F. Li, B. Luo, "Cyber-Physical Systems Security – A Survey," *IEEE Internet of Things Journal*, vol. 4, no. 6, 2017, doi: 10.1109/JIOT.2017.2703172.

[87] M. Stoicescu, J.-C. Fabre, M. Roy, "Architecting resilient computing systems: A component-based approach for adaptive fault tolerance," *Journal of Systems Architecture*, vol. 73, pp. 6–16, 2017, doi: 10.1016/j.sysarc.2016.12.005.

[88] M. Garcia, A. Bessani, I. Gashi, N. Neves, R. Obelheiro, "OS diversity for intrusion tolerance: Myth or reality?," in 2011 IEEE/IFIP 41st International Conference on Dependable Systems & Networks (DSN), 2011, pp. 383–394. doi: 10.1109/DSN.2011.5958251.

NoPASARAN: a Novel Platform for Analysing Semi-Active elements in Routes Across a Network

Ilies Benhabbour | King Abdullah University of Science and Technology in Thuwal, Saudi Arabia, ORCID: 0000-0002-0694-6767

Marc Dacier | King Abdullah University of Science and Technology in Thuwal, Saudi Arabia, ORCID: 0000-0003-3206-2030

Abstract

In this paper, we propose a novel, collaborative distributed platform to discover the presence, or analyse the configuration, of what we call semi-active elements. By doing so, we revisit the ideas initially proposed in [1, 2] with the Netalyzr tool and in [3] with Inmap-t. Our contributions lie in a simplified and more powerful design that enables the platform to be used for a variety of tasks, such as conformance verification, security testing, network configuration understanding, etc. The specifications, design and implementation choices of the platform are presented and discussed. Two use cases are revealed to illustrate how the platform can be used. We welcome any interest shown by others in deploying our tool in different environments, and encourage any subsequent collaboration in improving its expressiveness.

Keywords

conformance, firewall, `IPSEC`*, man-in-the-middle, network, proxy, security, TLS*

Corresponding author:
Ilies Benhabbour, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, ORCID: 0000-0002-0694-6767; ilies.benhabbour@kaust.edu.sa

Cite this article as: I. Benhabbour, M. Dacier, "NoPASARAN: a Novel Platform for Analysing Semi-Active elements in Routes Across a Network," ACIG, vol. 1, no. 1, pp. 89–118, 2022, DOI: 10.5604/01.3001.0016.1461

1. Introduction

Nmap is a well-known tool [4]. It enables, among other things, a client machine to fingerprint a remote one. We want to do the same thing for the devices on the path between two machines communicating together. In theory, this is useless because the devices on the route should never interfere with the application-layer connection. This is the well-known end-to-end principle proposed in the 1960s by Baran [5] and Davies et al. [6] and subsequently implemented in the TCP/IP architecture. The paper by Saltzer et al. [7] precisely articulates the arguments in favour of such a design for distributed systems.

These principles are certainly still valid. A TCP connection is, indeed, an end-to-end connection. However, the end of the connection is not always where the user expects it to be. For instance, imagine a user whose browser is configured to use a web proxy. When the user visits a web site, their view of the world tells them that they are connected to that web site. In reality, the TCP connection initiated by the browser ends at the machine where the proxy server resides. This is, of course, a trivial example. In today's networks, there are many other such instances where a device on a path between two communicating endpoints could interfere with their communication. Such interference can take various forms and take place at various layers. It can be as drastic as changing the TCP endpoint, such as in the case of the web proxy, but it can be more subtle, such as changing the source or destination IP (in the case of a NAT device) or delaying packets (in the case of a traffic shaper). As in [3], we collectively refer to all such devices as semi-active components.

Semi-active components aim at being beneficial for the end user. For instance, they can improve security (firewall, WAF, IPS, etc.) or performance (CDN, web proxy, etc.); they can also provide better access to the network (traffic shaper). Unfortunately, like the proverbial double-edged sword, their mere existence can also be detrimental to security. If compromised or misconfigured, the capabilities of such devices could be misused by attackers to their advantage. The same holds true if an intruder manages to insert their own semi-active component into a route between two devices.

One would hope that cryptography could come to our rescue to ensure true end-to-end connectivity. This is indeed the expected benefit when using TLS [8] (or IPSEC [9]) at the transport layer (or network layer). Once more, the devil lies in the details. The reality is that it is almost impossible for end users to be sure that, under the hood of the various layers, their connection is truly an end-to-end one. For instance, in many companies, a Web Access Firewall (WAF) will intercept, decrypt and re-encrypt TLS connections to protect end users, effectively deceiving them. From a networking point of view, there is almost no difference between a genuine WAF, installed by the right authorities, and a malicious device, inserted by an attacker, carrying a so-called man-in-the-middle attack, as described later.

As we can see, semi-active components matter. It is very important to verify that the legitimate ones behave the way they should, i.e. are not compromised or misconfigured. It is also very important, perhaps even more so, to detect the presence and identity of any other semi-active component

in the routes we use. The framework we propose in this paper answers the following two research questions:

1. How to detect and identify semi-active components between two endpoints?

2. How to verify that a known semi-active component is behaving the way it should?

To answer these questions, the paper is structured as follows. Section 2 positions our work with respect to state-of-the-art practices. Section 3 presents the architecture of our new platform. It starts by introducing some terminology (3.1), then clearly states its specifications (3.2), discusses its design choices (3.3) and ends with a description of its implementation (3.4). Section 4 proposes two simple use cases in order to exemplify the usefulness of the platform, its simplicity and modularity. Section 5 concludes the paper by mentioning a number of different use cases and by inviting readers to contribute to this collaborative platform, either by deploying one of its elements in their own networks or by contributing to one of its modules.

2. State-of-the-art Practices

Semi-active components have mostly been looked at in the literature through the prism of man-in-the-middle (MITM) attacks, in which an intruder intercepts a connection and impersonates one of the two communicating parties. Many variants of these attacks exist but the most frequently mentioned do take advantage of weaknesses in the ARP or the DNS protocols, as explained in [10]. Solutions have been proposed to detect these attacks. They are usually quite specific, focusing on the symptoms of a single type of threat. For instance, in [11] the authors propose relying on the delay introduced by the attacker during the processing of packets. This method, although intuitively correct, has a significant drawback. It requires knowing the ground truth value for each connection and thus it cannot easily be generalised to connections taking place on the Internet. In [12] the authors propose the use of a neural network to study the response behaviour and classify the connection. The authors of [13] note that deep-learning-based approaches are not robust to perturbations and that knowledgeable attackers may use adversarial attacks to bypass such detection. Finally, Trabelsi et al. propose a solution for detecting MITM attacks in the context of local area networks (LANs) [14]. Their approach is interesting in the sense that it also detects threatening devices by running some active tests. They do so in order to detect those who act as routers. It does rely on ARP cache poisoning though, which is possible in their target networking environment, LANs, but impractical on the general Internet.

In [15] the authors present a detection system that identifies web proxies. Their system records and analyses incoming packets to match them with well-known proxy patterns. To read the content of the encrypted payload and identify the patterns in the data, their method uses SSL stripping [16] to disable TLS. This method assumes that we can downgrade

a connection from HTTPS to HTTP and is thus impractical. Moreover, it also creates a security gap as all the traffic between their detection system and the end-user is now in clear text. Chiapponi et al. present in [17] a detection method aimed at bots using proxies to scrape web sites. This method is based on the round-trip time difference for the packets sent from the server to the proxy and those sent to the bot itself (through a TLS tunnel). A large measurement campaign validates this idea experimentally, and the data are then used in [18] by Champion et al. to find a method capable of geolocating the machines involved.

Our new contribution has mostly been inspired by two other pieces of work: Netalyzr, by Kreibich et al., published a dozen of years ago [1, 2], and the more recent work by Vitale et al. [3] on Inmap-t.

As far as we can tell, Netalyzr [1] represents the seminal work in the detection of semi-active components. It is based on a client-server architecture. A user connects to the server and downloads a Java program which runs within their browser. This program executes a series of tests against the connection, with the server aiming at detecting the presence of semi-active components between the client and the server. For instance, in [2] the authors use it to discover the presence of a proxy server between a client and their server. In their tests, they were sending well-crafted packets to trigger some side effects induced by semi-active components. Since both the server and the Java program are synchronised, they know what is supposed to be sent at any point in time. By comparing what they receive with what they expect to receive, they can infer the presence of some semi-active components. For example, in one test of [2], the server replies to the initial SYN packet of the establishment of a 3-way TCP handshake, by a TCP RST packet as opposed to the expected SYN–ACK one. If no proxy exists between the client and the server, the Java program will indeed receive the sent RST packet. If a proxy exists, its kernel will, in most cases, complete the 3-way TCP handshake with the client before receiving the RST packet from the server. The Java program infers the existence of a proxy as soon as it receives the unexpected SYN–ACK packet. This solution has been available as a free service for several years but has unfortunately been discontinued since 2019. Maintenance costs of the service and the issues associated with the Java language are the reason for its demise [19]. Although Netalyzr represented a laudable first attempt at detecting semi-active components, it was only capable of detecting them on paths that were leading to the targeted test server. A knowledgeable attacker would therefore have had little difficulty in circumventing the detection provided by this solution.

Inmap-t, proposed in 2021 [3], aimed at leveraging the Testing and Test Control Notation Version 3 standard (TTCN–3) [20] to test the security impact of intra-network elements. TTCN–3 is a standard maintained by ETSI that offers a modular testing language and an independent execution environment. Its usage is normally reserved for testing the quality and conformance of a given implementation of a client or server for a specific protocol. In their work, the authors instead leverage the environment to test the network connection taking place between two machines, namely, to detect the presence of semi-active components such as an Intrusion Prevention System (IPS) [21] or a firewall. As opposed to Netalyzr, the

authors now offer a fully distributed environment populated with numerous devices that can communicate together and test a number of different paths. The results, while promising, also reveal the drawbacks associated with the choice of `TTCN-3` as an underlying platform. According to the authors, the learning curve to use this environment is very steep and the heavily C++-inspired notation of the configuration files does not help in that regard. Furthermore, `TTCN-3` comes with its own execution environment whose installation is quite heavy and cumbersome. Last but not least, the software architecture design imposes the constraint of grouping all possible tests into a single binary. The net result is that every new test increases the overall size of the code to be pushed on all participating machines. Every modification to an existing test requires pushing a whole new version of all tests to all machines.

Netalyzr and Inmap-t have shown the value and the feasibility of detecting semi-active components. They also have limitations that hold back their wide adoption for carrying out these tests systematically. In this work, we leverage the lessons learned from these previous attempts and come up with an architectural design, which not only addresses their shortcomings but also greatly increases the diversity of tests that we are able to run.

3. Architecture

3.1 Terminology

Our work leverages the key concepts developed in [3], but greatly simplifies the architecture and, more importantly, enriches its capabilities with novel contributions.

In [3] the authors had to introduce the `TTCN-3` terminology for their system to be understandable. It included several well-defined terms, such as *Main Test Component (MTC)*, *Parallel Test Component (PTC)*, *Test System Interface (TSI)*, *Port, Module and Verdict*. Although we are not using any of these notions, we retain two key ideas: i) *test case* and ii) *test campaign*, and we redefine them as follows:

Test case: A test case is a small program running in a synchronized way on two (or more) machines to test a specific property of an element in a given route. It defines the details of the sequence of packets to be sent, as well as the information exchanged between the machines while running the test case. It is typically defined by a finite state machine.

Test campaign: A *test campaign* is a larger program running in a synchronized way on two (or more) machines to characterize one or more elements in a given route. It can be seen as a sequence of test cases or of other *test campaigns*. It is typically defined by a decision tree whose main components are test cases or *test campaigns*.

Primitive: A primitive is a simple function that we define. Its implementation remains hidden from the user and usually involves some Python and Scapy

[22] code. Test cases can only be made up of primitives. This enables us to completely separate the definition of a test case from its implementation. The creator of a test case is not going to be exposed to Python code. Moreover, if we decide to change from a Python implementation to a C++ one, for instance, the test cases and test campaigns would remain untouched.

3.2 Specifications

In this section we describe the high-level specifications of what we aim to achieve in using our new solution. The next subsections will cover our design and implementation choices (subsections 3.3 and 3.4).

At the highest level, what we want to do is to build a system that enables us to answer the two research questions defined in section 1, namely:

1. How to detect and identify semi-active components between two endpoints?
2. How to verify that a known semi-active component is behaving the way it should?

Building upon the lessons, drawbacks and weaknesses from previous work [1, 3], we want our solution to satisfy the following five properties:

1. Ease of use
2. Ease of deployment
3. Modularity
4. Scalability
5. Flexibility

We now briefly elaborate on each of these properties.

Ease of use: We target end users who do not have any particular networking or security knowledge. The solution must hide all its low-level complexities and provide a simple interface for them, such as verifying whether or not a `WAF` monitors their connection to the Internet.

Ease of deployment: Our platform aims at being a large, open and collaborative distributed system. The more people deploy our solution at their site, the richer the system becomes. This precludes a complicated time-consuming system setup, such as the one required by a `TTCN-3` based solution.

Modularity: The creation of a new *test campaign* comes down to combining previously defined test cases and/or *test campaigns*. The same holds true for test cases that can take advantage of previously defined finite state machines by "calling" them in. Our software environment must facilitate such code reuse of test cases and campaigns by defining them as well-defined modules, with precisely specified input and output.

Scalability: As opposed to the solution described in [3], the definition of a new test case should not increase the code size of all existing *test*

campaigns. Also, the cost and complexity for distant machines to run a given *test campaign* should be independent of the number of machines participating on our platform and running other *test campaigns* independently on their own.

Flexibility: As opposed to the solution described in [1], we do not wish to limit our analysis to the routes that lead to a single server providing the Java code. We also want to avoid being confined to the sole routes connecting the machines participating on the platform, as in [3]. We want to be able to test the properties of the route between any of the end user machines and any server on the Internet.

In the next subsection, we outline the design choices we have made to build an open platform that would satisfy these properties.

3.3 Design

What we are trying to achieve can be summarized as follows:

- Some remote parties decide to test the properties of a specific route on the Internet.
- They agree on what data to send: where, when and how. This is defined in the *test campaign* that they all agree to run.
- To evaluate a given property, they compare the packets they receive to the ones they expect to receive, as per the definition of the *test campaign*.
- They adjudicate on the results of the tests.

To justify our design choices, we present them according to three major components of the platform we have built, namely:

1. Overall architecture
2. Data and Control communication channels
3. Test cases and *test campaigns*

Architecture: Our architecture consists of three types of machines: **worker**, **proxy** and **master**.

A worker is a machine that runs a *test campaign* – either the client machine that wants to perform a test on its connection path to another endpoint or a trusted machine registered in the network. A proxy node, on the other hand, does not perform any test. It should be accessible to the remote workers so that they can communicate together when, for instance, they are unreachable from the Internet because of a firewall.

The master node role puts the workers (and the proxy when necessary) in touch and shares with them the *test campaign* they have to run. Once this is done, the workers do not use the master node anymore[1].

1 —— This is a major difference with the architecture proposed in [24], in which all communications had to go through some central component. This represents a bottleneck when many workers are running tests at the same time.

Test cases and test campaigns: Test cases are defined thanks to a graphical user interface using a finite state machine formalism. They indicate by means of well-defined primitives what packet should be sent by whom to where, what to do when receiving a packet, what field to check, etc. These finite state machines can then be combined into *test campaign*s, under the form of decision trees. *Test campaign*s are stored within the master node and pushed to workers when they need to be run. All workers receive the same *test campaign* and interpret it according to the role each has to perform. To synchronise their execution, the workers do exchange control messages using the channels described hereafter.

Data and control communication channels: We use two channels to enable the workers to execute a *test campaign*: i) a data channel and ii) a control channel. The data channel is used to send packets on the route we want to test. The control channel, as its name implies, is used to exchange control messages between the workers to synchronise the execution of the tests. It is worth noting that the data channel does not necessarily follow the same route as the control channel.

At this stage, it is probably worth providing a simplified high-level example of what our solution aims at doing and how this can be achieved.

Let us imagine that we have two workers w_1 and W_2, a proxy `p`, a master node `m` and a `DNS` server `d`. w_1 wants to know whether its `DNS` requests to `D` are intercepted by a third party and redirected to another `DNS` server that would provide a different `IP` for a given request `r` than the one `D` would return. This is one of the classical ways to redirect web request to a `WAF` without having to touch the user's machine configuration. Let us further assume that both w_1 and w_2 are located behind firewalls and cannot be contacted directly by one another. This situation is represented in Fig. 1.

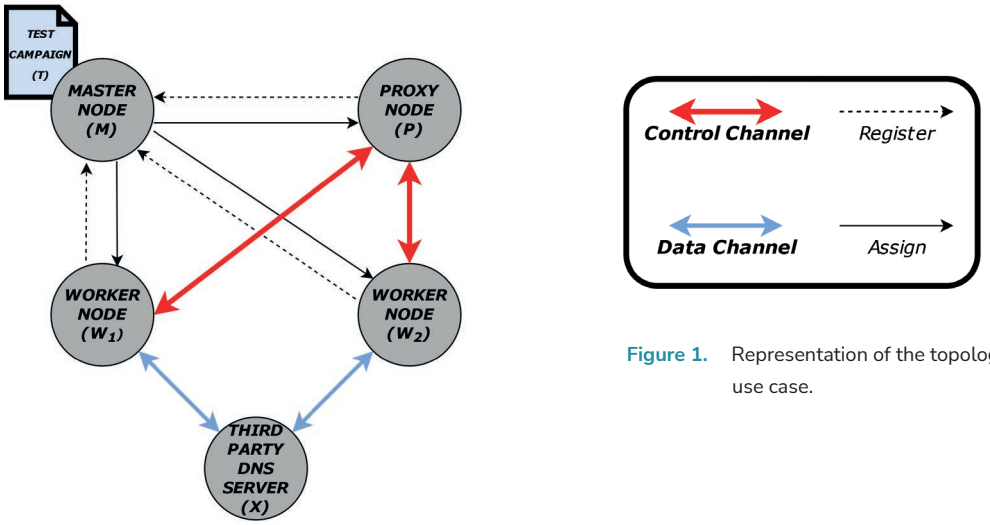


Figure 1. Representation of the topology for our use case.

To run this test, the following steps must be performed:

1. w_1, w_2 and p register with m and maintain a secure connection to it.
2. w_1 issues a request to m to run the test campaign t with w_2.
3. M assigns P to act as a proxy for the control channel between w_1 and W_2.
4. Both w_1 and W_2 establish an IPSEC tunnel to P and use them to establish a trustworthy end-to-end TLS connection between w_1 and W_2, passing through P. In the latter, we will refer to that secure end-to-end connection as the control channel.
5. m uses the existing secure connections to push T to w_1 and w_2.
6. w_1 and w_2 send r to d.
7. W_2 sends the result of the DNS answers to w_1 through the control channel.
8. w_1 compares what it received from W_2 with its own DNS replies and reaches a conclusion for the test accordingly[2].

2 —— This is a simplified view of the world. We are well aware of the possibility, in some cases, to obtain distinct responses from a given server d for the same request r issued by distinct clients. The identification of the reasons for this to happen could, in fact, be an interesting use case for our platform but such a discussion lies outside the scope of this paper.

This very simple example highlights a couple of important elements. First, the route to test is not necessarily between the nodes we own. Second, the control channel is not necessarily a direct connection between two workers, it can go through a proxy when needed. Third, the campaigns are uploaded on a case-by-case by the master node to the workers. Fourth, the campaigns are functions whose input parameters are instantiated by the workers and agreed upon via the control channel.

The next subsection gives more details on the way this platform has been implemented.

3.4 Implementation

The previous subsection presented the design of our solution based on three elements: the tests, the communication channels and the architecture. In the following subsection, we describe the implementation of each of these.

Test case: We represent a test case using finite state machines. The four main components that constitute these state machines are the following:

- State
- Transition
- Action
- Nesting feature

To present each of these components we use the simple example presented in Fig. 2. We present two simple, yet genuine, use cases in section 4.

State: A state is represented by a light grey box with a blue header in Fig. 2. It is defined as a sequence of actions. Every finite state machine starts with an initial state that triggers an event started at the beginning. A state possesses its own local variables. To transition from one state to another, we need to trigger an event whose name matches a transition that links these two states.

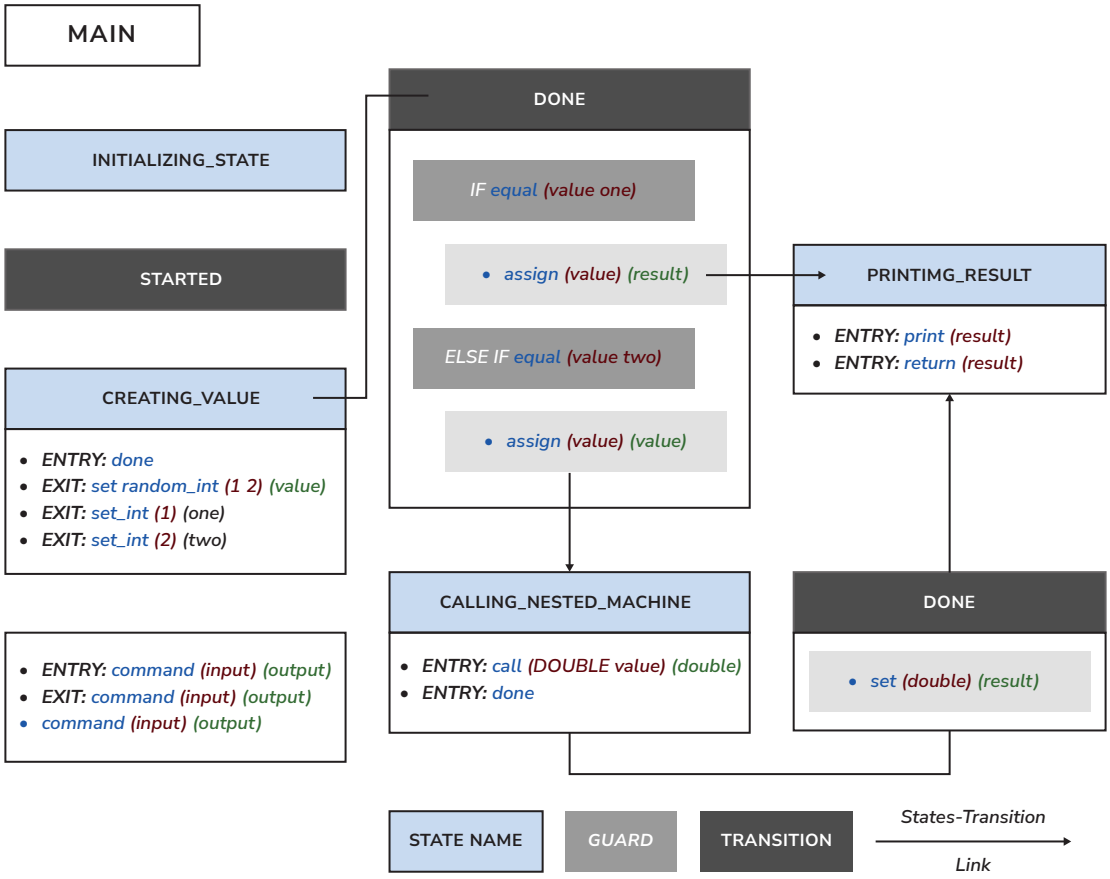
Transition: A transition is represented by a light grey box with a dark grey header (Fig. 2). It represents a direct path from one state to another. To leave the current state, an **event** with the same name as a transition linked to that state must be triggered. Depending on its implementation, the execution of a primitive can directly generate events responsible for this change of state. This is, for example, the case when we call the `done` primitive to trigger the `done` event. Events can also be generated by external elements, such as the reception of a packet or a timeout. A transition may use **guards** to specify the conditions on the local parameters of the current state.

Action: We refer to a command that a state can execute as a **state action**. It can fall into two categories, depending on the moment we want the program to execute it:

- An **Entry** action executes a primitive as soon as we enter a new state.
- An **Exit** action executes a primitive when one of the possible events to leave the current state has been caught (but after having executed all the *Entry* actions).

Since states possess their own set of local variables, we define a **transition action** as the special action that uses the `set` primitive to assign the value from one state local variable to another state local variable.

Nesting feature: To create modular, understandable and easy to modify test cases, we implement something referred to as the nesting feature. This mechanism enables us to call nested finite state machines. We carry out this operation with the `call` primitive in the parent. Then we use the `get_parameters` primitive to retrieve the input arguments in the nested state machine. The number of parameters associated with this primitive must match the number of parameters when the parent uses `call` (without the name of the state machine called). This is shown in Fig. 2, where the `main` state machine calls the `double` one.



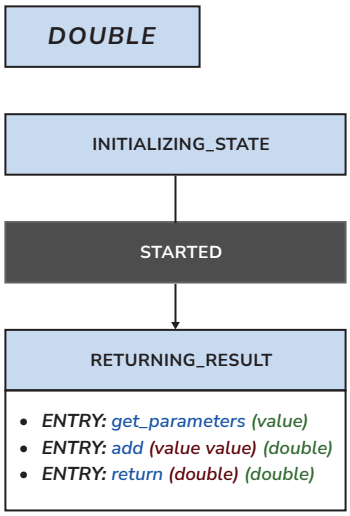


Figure 2 Representation of our implementation of two finite state machines. The first one (top) is responsible for selecting a random value between 1 and 2. If the number is 2, it doubles this value by calling the second one (bottom). At the end, the main finite state machine prints the result and returns it back.

> Similarly, the number of output parameters of the `call` primitive must match the number of parameters provided by the `return` primitive. The idea behind this is to create a higher level of abstraction that is similar to traditional programming libraries. We take advantage of other finite state machines already written, so that we can reuse them across different test cases.

The transformation from a state machine representation to executable code is a key functionality that our solution implements. We use Xstate [23] as it provides a complete set of features to easily manage state machines, such as adding states, actions or transitions. Its main advantage is that we can export the representation of the state machine into `JSON` format afterwards. Our Python program then parses this `JSON` to automatically obtain the program to run for the test case.

Test campaign: As we mentioned earlier, test campaigns are composed of test cases. When test cases end, they can output different values using the `RETURN` primitive followed by the returned values. We use decision trees to aggregate their results and define test campaigns. Let us imagine that we want to do a test campaign that runs the `MAIN` test case two times consecutively and we want to know the sum of both results. The test case can only return 1 or 4. The decision tree representing the test campaign is shown in Fig. 3. We start from the root and each node represents a test case where its return value gives us the next node in the path. The test campaign finishes when it reaches a leaf node. The possible results are thus 2, 5 or 8.

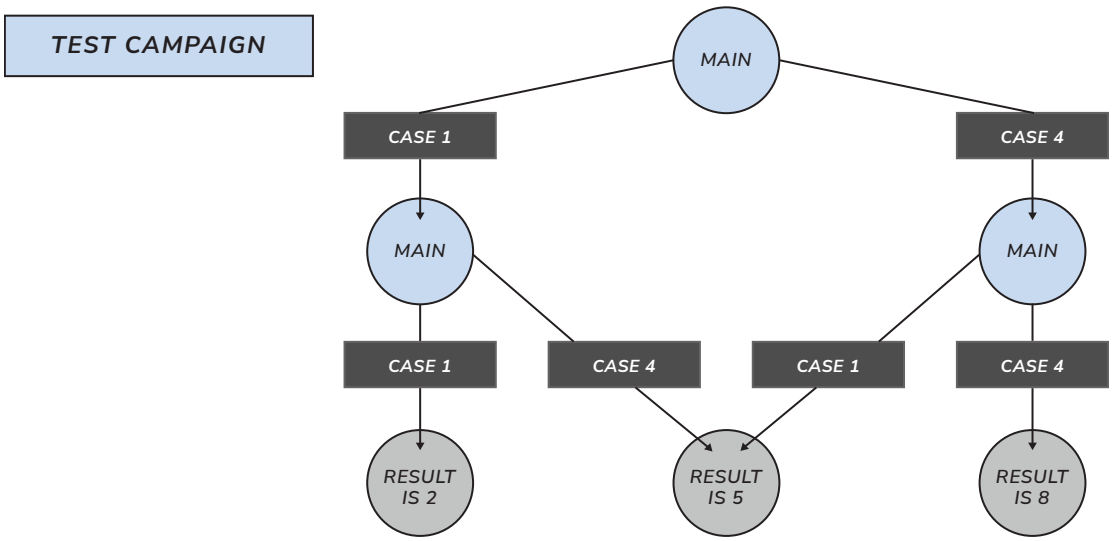


Figure 3. A decision tree that repeats sequential test cases until it reaches a leaf node where it outputs its result.

In that case the example is trivial, yet it shows how to combine test cases into a *test campaign* to answer questions that are more complex. We can reuse unitary tests for different *test campaigns*. This hides the complexity from the user and it provides us with a modular solution.

Data channel: The data channel enables us to send (or receive) packets to remote machines that we do not necessarily control through the `send` (or `wait_packet_signal`) primitives. We emit, filter and sniff packets using the Scapy library in Python. Each transmission of a packet `triggers` a `packet_sent` event. Received packets that passed the filter phase are stored in a `FIFO` data channel queue to ensure that no packets are lost. When the program executes the `wait_packet_signal` primitive, it checks whether packets are available inside the queue whose name was given as argument. If this is the case, a `packet_available` event is generated and the oldest packet in the queue is processed. Otherwise, after a certain time has elapsed, it raises a `timeout` event.

Control channel: We specify the behaviour of the control channel in the definition of the state machine to synchronise the workers. This comes with its dedicated set of primitives and events to maintain the correct flow of the state machines.

- The primitive `wait_ready_signal` triggers the event `READY` when the node successfully creates a control channel.

- The primitive `wait_sync_signal` triggers the event `sync_available` when the node has an available sync message in the control channel queue.

- The primitive `sync` triggers the event `sync_sent` when the node sends a sync message.

  We define a **sync message** as a message whose reception acts as a checkpoint. It is responsible for the synchronisation of the worker nodes. A sync message can embed data to provide a remote node with some information securely using the control channel. We use the same queuing mechanism as the data channel to listen constantly for sync messages. This guarantees that no packet is lost due to concurrency issues between worker nodes. We call it a **control channel queue** and it is unique for each control channel created between two nodes. Section 4 provides two use cases to show how our framework takes advantage of it. Finally, to ensure that we can start sending packets through the control link, we use the `wait _ ready _ signal` primitive. As soon as the channel is ready, it triggers the event `READY` to resume the execution of the test case.
  In order to protect our control channel, we use `TLS` and `IPSEC` where necessary. Our solution uses `TLS` with mutual authentication to link two remote worker nodes. We encapsulate `TLS` using `IPSEC` to create a tunnel that connects the worker nodes and a proxy when they are hidden behind firewalls. It enables the workers to redirect the traffic so that they can initialise the control channel. To ensure that all nodes perform proper mutual authentication, we designate a master node as the only trusted authority. This is the sole node able to sign the certificates associated with the different public keys announced by the other nodes. This architecture enables us to revoke the certificates of compromised machines using the `OCSP` [24] protocol.

Architecture: In order to assign the test campaigns to the different nodes of the network, we use Ansible [25]. Its use only requires Python and `SSH`, which are often installed by default on Linux machines. The description of that part resides outside the scope of this paper.

In the following Section, we provide two simple use cases, one for `UDP` traffic and the other for `TCP` connections, to exemplify how these concepts can be put into action from a practical point of view.

4. Use cases

4.1 DNS redirection

Our first use case is the one eluded to in subsection 3.3 in which a client tries to determine if their DNS server returns different results than the ones other users are seeing for the same domain name. There are many reasons why network administrators routinely do this. To render things concrete, we can think of a network that uses a Web Access Firewall to check any outgoing HTTP request in order to detect those coming from possibly compromised machines trying to "*call home*", or to block requests sent to sites forbidden by the local security policy. This is a very common practice in enterprise networks. To force the traffic to go through the WAF, several techniques exist, such as using DNAT or WPAD. Another simple technique consists of using DNS. Whenever a client uses DNS to resolve the name of a web server it wants to access to, the returned IP is that of the WAF instead of the actual machine. The client initiates the connection to the IP of the WAF, which decides whether the request is to be blocked or not. If the traffic is allowed to go through, the "Host:" HTTP header enables the WAF to forward the traffic to the right destination. There is a problem if the initial DNS request was not made to generate some HTTP traffic afterwards. If this is indeed the case, this approach prevents the establishment of the connection, since the WAF does not find in the application payload the identity of the server to contact on behalf of the client. Consequently, using this technique requires using heuristics at the DNS server level to decide whether or not to return the real IP or the IP of the WAF to avoid blocking non-HTTP traffic. Domain names starting with "www" will typically be resolved by the WAF IP, whereas names starting with "ftp." or "smtp." will not, for instance.

In our use case, a client wants to determine whether or not their network applies such DNS-based protection and, if so, which heuristics are being used at the DNS server level.

Setup: We have a client w_1 who wants to detect if DNS requests sent to D are redirected using another remote machine W_2. Both workers w_1 and W_2 have registered their availability to the master node m and maintain a secure connection to it. Upon w_1's request, m notifies W_2 that w_1 wants to run a test campaign with it. If W_2 agrees, M, which knows that both are sitting behind a firewall, assigns a proxy p and provides the test campaign T along with its parameters to all the parties. w_1 and W_2 establish an IPSEC tunnel to p and then a TLS connection between them through p. They finally instantiate two separate data channels with D. We use Fig. 4 to illustrate our architecture. The client can then repeat this process with different requests to identify the heuristics used, if any, or with other w_k workers to double check its results.

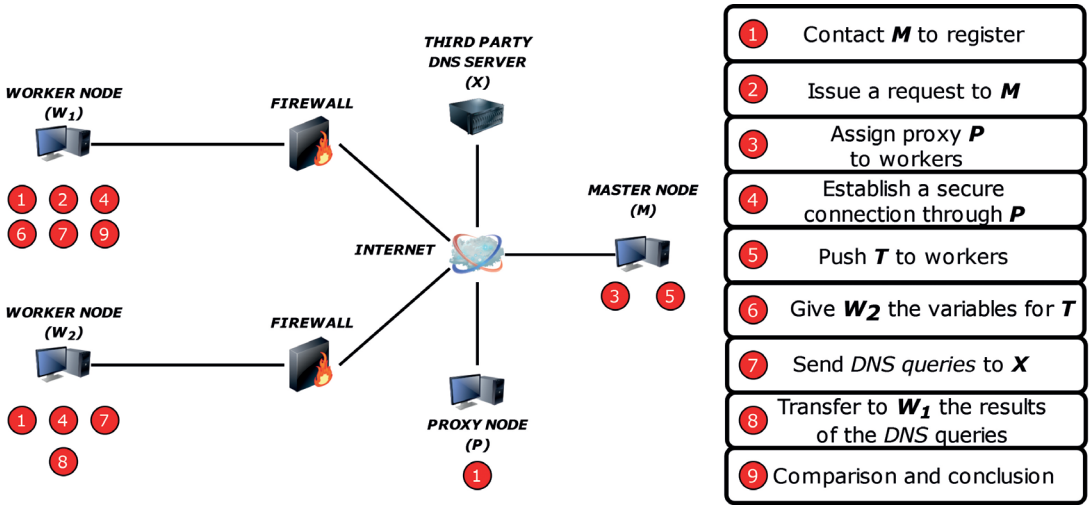


Figure 4. Architecture used for the DNS use case.

Test case: We present here how this use case can be modelled as a finite state machine. As mentioned in subsection 3.4, we rely on the Xstate graphical interface [23] to create the state machines and to automatically generate a JSON file that contains all their semantics. That file is then sent to both workers, parsed, and interpreted by a generic piece of code capable of executing any such state machine. It is worth noting that, even though both workers execute different tasks, they both interpret the same file. We have made this design choice to avoid having numerous files associated with a given use case that could, if modified independently, become out of sync. Having only one file per use case for all workers eliminates this risk.

Main finite state machine: This test case can be represented by only 7 states, which are represented in Fig. 5. It is thus easy to read and understand what this use case represents. The simplicity is obtained thanks to the modularity of the design. Indeed, each state in this finite state machine corresponds to a full finite state machine that remains hidden at this level of abstraction. We will briefly list the various states and their functions. We discuss each state in more details afterwards.

`defining_timeout_target`: a catch all state to handle all timeout events in a generic way.
`control_channel_set_up`: builds the control channel.
`variables_set_up`: builds the DNS packet according to the input provided by w_1 to m and returns the name of the data channel queue on which the received packets are stored.
`dns_request`: sends the DNS request and waits for the reply.
`remote_reply_getter`: exchanges the DNS replies between the worker nodes.
`dns_replies_comparison`: compares the DNS replies.

`control_channel_disconnect`: closes the control channel.

In what follows, we will briefly describe all the state machines required to run this use case, starting from the main one at the highest level of abstraction, and following by the ones that are invoked from within this main one.

MAIN: This state machine is the main component of our test case. Its execution is triggered by `M`. It runs on w_1 and W_2.
Like every state machine, its initial state is named as `initialising`. Launching the execution of the finite state machine (or invoking a finite state machine, in the case of a nested feature) generates the event `started` which triggers the first transition of the state machine.
In this state machine, all subsequent states will invoke the `call` primitive to execute another state machine.
The end of the state machine is characterised by a state named `ending`, which invokes the `return` primitive with a list of arguments that represent the results of the execution of the state machine.

DEFINING_TIMEOUT_TARGET: This first state is very specific. It only invokes the primitive `redirect` before generating the done event to move to the next state. The primitive `redirect` associates an event with a specific state. In this case, we associate the `timeout` event with the `error` state. This means that, no matter where we are in the finite state machine, if the `timeout` event is raised then the next state of the state machine will be the `error` state. This is a convenient and simple error handling mechanism.

CONTROL_CHANNEL_SET_UP: This state uses the `call` primitive to invoke another state machine, the role of which is to create the control channel between w_1 and W_2.
Whether this control channel goes through a proxy or not is irrelevant for the use case and remains hidden to the creator (or user) of the state machine. The usage of a proxy will be required if both are hidden behind a firewall. If only one is unreachable, that one will be the one initiating the connection to the other one. If both are reachable, `m` decides which one initiates the connection.

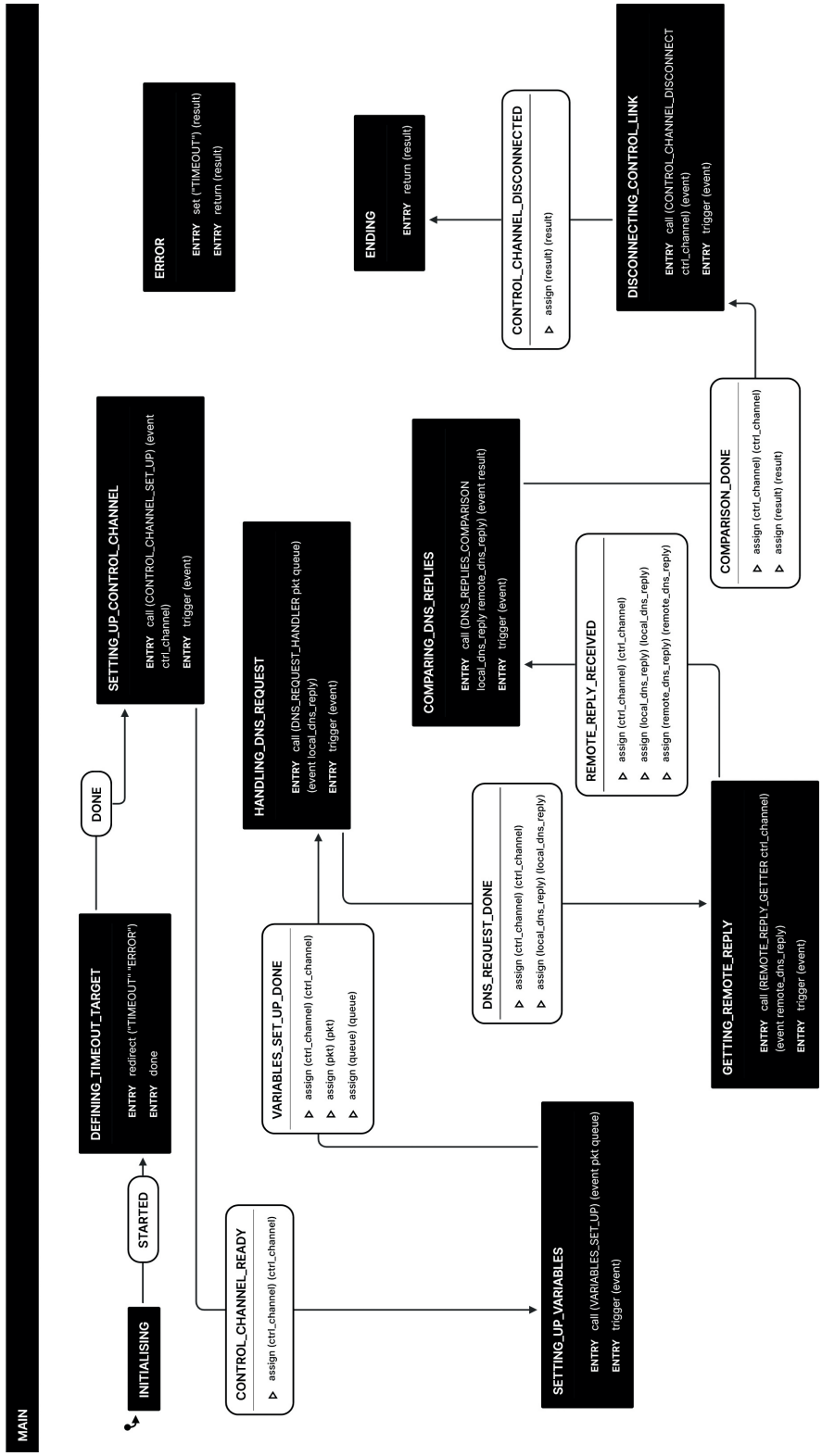


Figure 5. Main state machine for the DNS `redirection` test case.

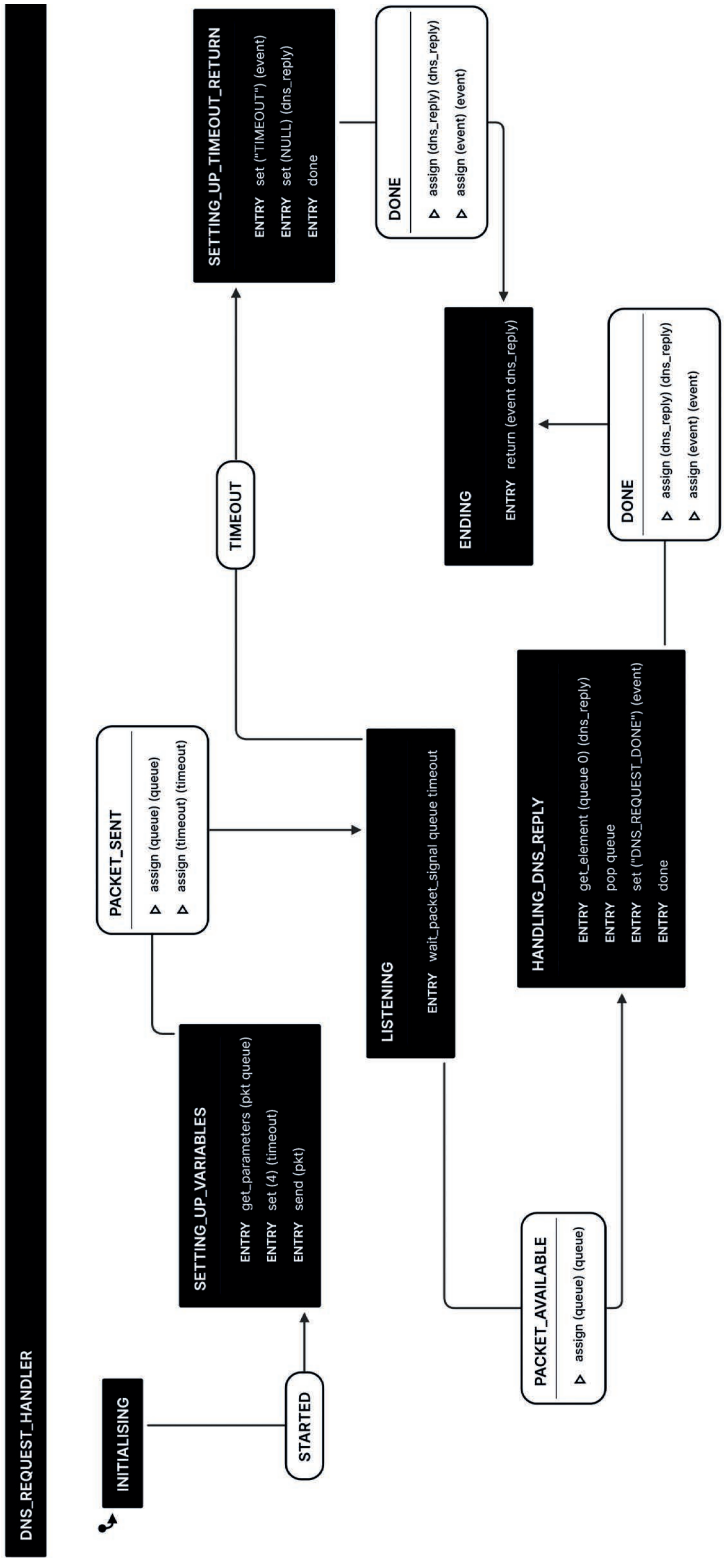


Figure 6. **Nested state machine that sends and receives DNS packets.**

Generally speaking, it is `m` that, together with the test campaign, informs the workers (and the proxy when needed) as to whether they have to initiate a connection (and to whom), or whether they have to be ready to accept a connection (and from whom).
When the invoked state machine has terminated, the `trigger` primitive is invoked. As its name implies, it triggers an event in order to move to the next state. The specific event triggered is the output of the called state machine. In this case, there is only one possibility, `control_link_ready`, but in the general case there could be several results leading to different states.

VARIABLES_SET_UP: This state is responsible for creating the same `DNS` request `r` on w_1 and W_2. All the low-level complexity is hidden behind the nested state machine. It will, for instance, create a `UDP` packet with destination port 53 corresponding to the `DNS` protocol while the source port is set as random. It also specifies the destination `IP` address and the requested domain name, as specified by w_1 to `m`.
In addition to the packet, the state machine defines a filter on the source port of the packet for the sniffer, so that the workers only listen to the `DNS` replies sent by `D` for the request `r`. The nested state machine links this filter to a data channel queue before starting to listen and push packets inside it.
At the end, the state forwards the name of the queue and the crafted packet to the next state.

DNS_REQUEST: This state calls a nested state machine responsible for the emission of the crafted packets and the reception of the `DNS` replies on both w_1 and W_2. Its representation is shown in Fig. 6. The two worker nodes send the `DNS` queries to d using the send primitive. They process the `DNS` replies stored on the data channel queue using its name and the `wait_packet_signal` primitive. The state then forwards the `DNS` replies received to the next state to compare if they are the same on both workers.

REMOTE_REPLY_GETTER: Before disconnecting the control channel, we need to ensure that the `DNS` replies received by w_1 and W_2 are the same.
We use the **sync** mechanism presented in subsection 3.4 to fully synchronise our state machines. The following steps show how it works in that specific case.

1. A worker node uses the `SYNC` primitive to send the payload of the `DNS` reply received within a sync message to the other worker.

2. It switches to a new state using the `sync_sent` event that is triggered and waits for an available sync message from the other node using the `wait_sync_signal` primitive.

3. When a sync message is available on the control channel queue, the state machine triggers a `sync_available` event to leave

the current state. The node finally extracts and stores the payload inside a variable. That variable is returned to the main state machine.

If an error occurs during the exchange and one of the nodes does not receive the sync message, the program triggers a `timeout` event. Similarly to try and catch exception handling in standard programming language, this will be given back to the main state machine to generate the final error output.

DNS_REPLIES_COMPARISON: In this state w_1 and W_2 compare the payload of the two `DNS` replies. If the `IP` addresses are the same, we set the result of the test as `EQUAL`. Otherwise it is set as `DIFFERENT`.

CONTROL_CHANNEL_DISCONNECT: Finally, we use this state to cleanly disconnect w_1 and W_2. Once again, we take advantage of the **sync** mechanism to perform this operation. In this case, w_1 and W_2 send a sync message when they are ready to disconnect. When they receive the remote sync message, the workers can terminate the control channel connection. This operation is necessary as the execution of the program on each node is concurrent, and we do not want one to disconnect while the other is still running some computation.

In our case study we have described how our solution is modular and simple. Indeed, we assign each specific task to an independent state machine. We can also easily modify the flow of the test case by removing actions, `redirect`ing transitions to other states or calling entire state machines. In addition, the use case shows how easy it is to create, send and listen for packets using the primitives in our toolbox. In the next subsection, we show how we can take advantage of a data channel between two worker nodes to infer the presence of a transparent web proxy and discriminate between the various types that exist.

4.2 TCP proxy

To further assess the effectiveness of our solution, we use it to create a state machine that can classify a proxy acting as a man-in-the-middle. The proxies we study in this use case terminate `TCP` sessions between a client and a server without any configuration on the client. As they are invisible to the client, we refer to them as transparent proxies. They pretend to be the server to the client in question for each `TCP` session and create another session with the real server to deliver the correct content to the client.

Proxy categories: To classify a transparent web proxy we can take advantage of the fact that its retransmission behaviour can fall into three categories. As represented in Fig. 7, a transparent proxy can either be:

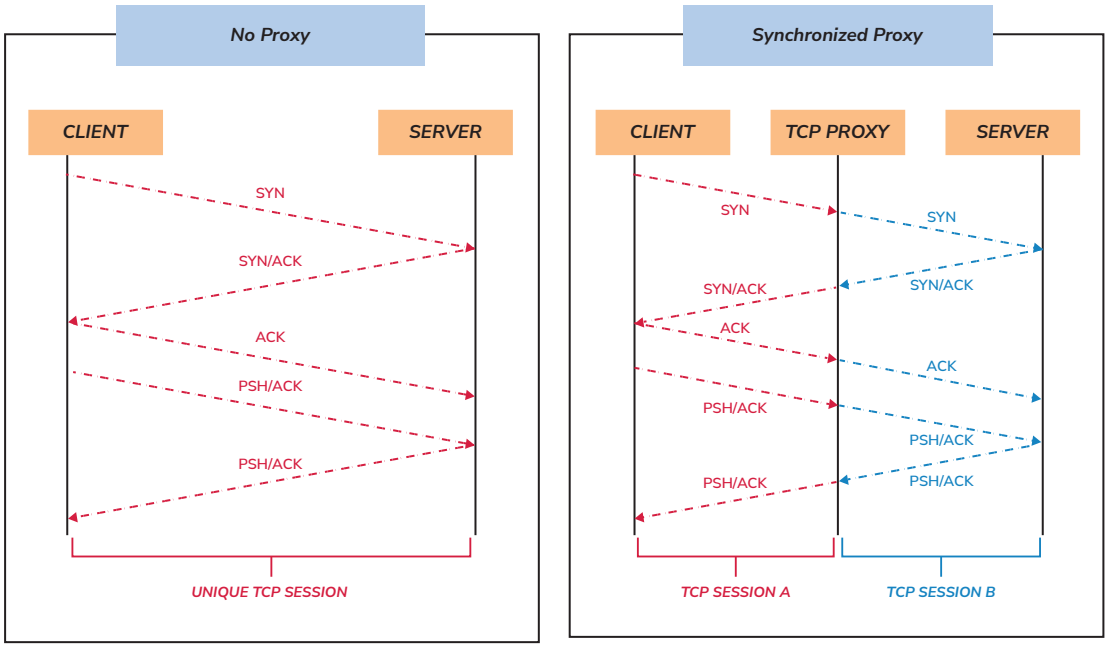
- **Synchronous**, when it only processes and forwards the packets which are received with a `FIFO` approach.

- **Asynchronous**, when it immediately responds to the SYN packet sent by the client with a SYN–ACK and forwards the SYN packet to the server simultaneously without waiting for an ACK.

- **Sequential**, when it does not initialise any TCP session with the server before it receives a PSH–ACK packet from the client.

Test case: To verify if there is a web transparent proxy in the communication path, and its category, we define the following test case with two worker nodes w_1 and W2 where they respectively play the role of the **client** and the **server**:

1. w_1 and W_2 create a control channel between each other.
2. w_1 sends a SYN packet on port 80 to W_2 while we prevent W_2 from sending the SYN–ACK reply.
3. w_1 (or W_2) listens for incoming SYN–ACK (or SYN) packets.
4. w_1 and W_2 send **sync messages** to each other to ascertain whether the other party received the expected SYN or SYN–ACK packet. These messages also embed the initial sequence number of the connection (set to NULL if W_2 did not receive the initial SYN.
5. w_1 and W_2 notify each other with sync messages that they are ready to disconnect from the control channel.
6. w_1 and W_2 terminate the control channel connection.



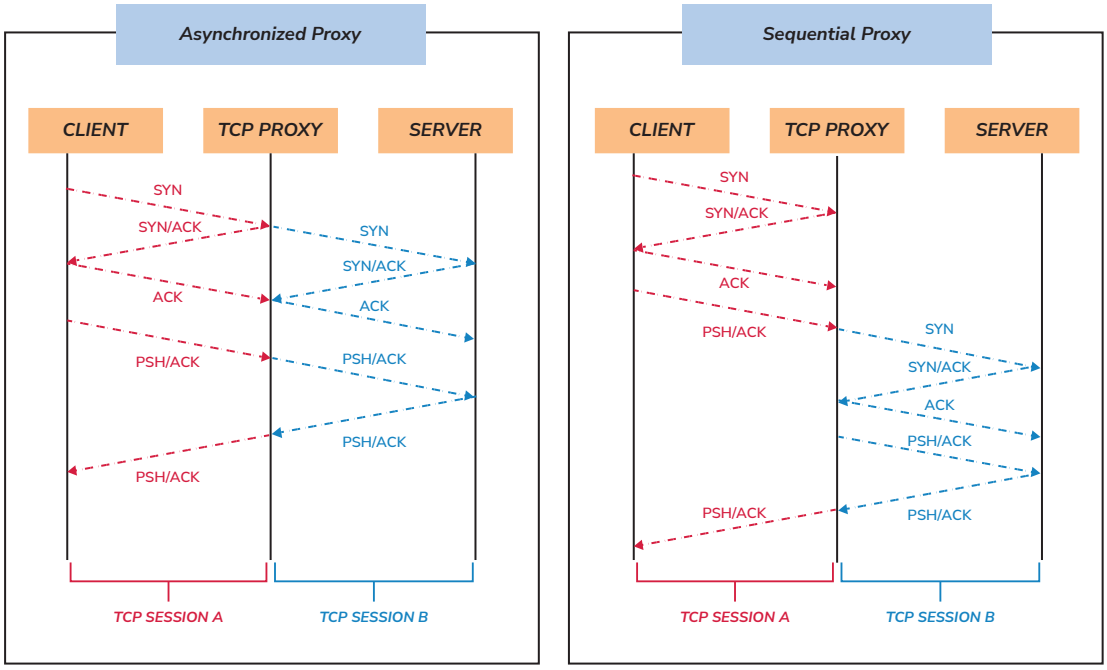


Figure 7. Representation of the behaviour of the connection with and without the three different TCP proxies.

Fig. 8 illustrates how the exchanges of packets on the data and control channels work. After it receives the information from the remote node, a worker can reason about the presence and category of a transparent proxy. For instance, if W_2 did not receive any SYN before a certain time, the test is over since we know that there is a **sequential** proxy[3]. If it received it, we check whether w_1 received a SYN-ACK packet. If so, we conclude the presence of an **asynchronous** proxy. Finally, if no SYN-ACK packet has been received by w_1, we compare the sequence numbers of the initial SYN packet for the TCP connection. If they are different, there is a **synchronous** proxy between the two nodes. Otherwise, we conclude that there is **no** proxy.

3 —— To simplify, we take for granted that the establishment of an HTTP connection between w1 and w2 (through a transparent proxy) would have been successful.

Main finite state machine: The main state machine for the aforementioned test case is represented in Fig. 9. Again, it is simple to understand as all the low-level primitives are hidden inside the nested state machines. It also shows how we can take advantage of the nesting feature to use nested state machines from other use cases. For instance, this test case uses three states already defined and presented in subsection 4.1.
In addition to these states, we present four new states dedicated to inferring the presence and category of a transparent web proxy:

`variables_set_up`: builds the SYN packet and provides the name of the data channel queue on which the packets will be stored.
`syn_handler`: w_1 sends the SYN packet, both w_1 and W_2 wait for a packet on their data channel queue.
`remote_received_getter`: sends to the other worker whether it received the expected TCP packet and the initial SYN sequence number.

`result_setter`: gives a conclusion based on the packets received and the `SYN` sequence numbers.

As we already described three of the state machines in subsection 4.1, and since they are the same in this use case, we only present the new states in more detail.

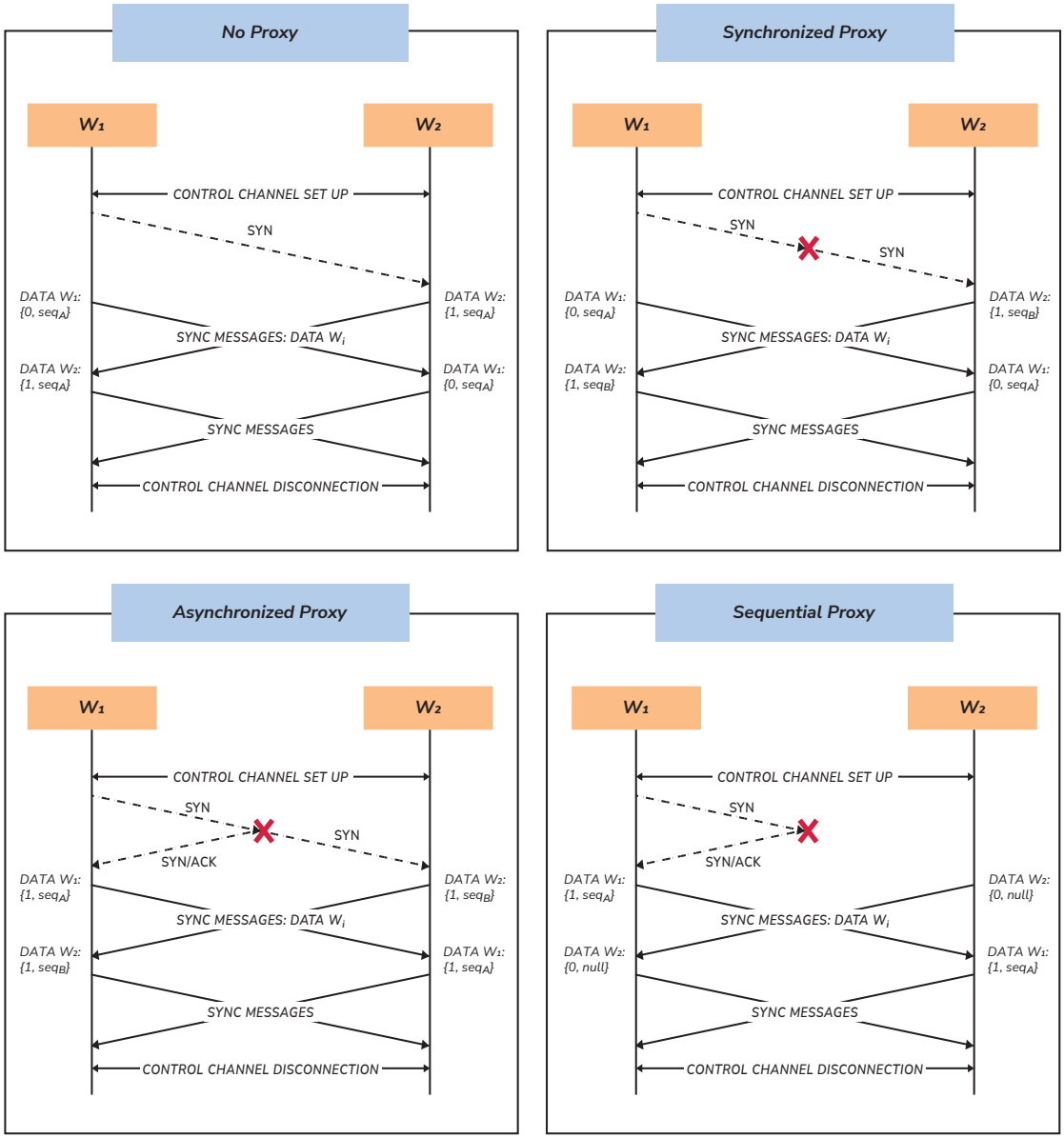


Figure 8. Representation of the four possible scenarios, depending on the presence and category of a `TCP` proxy between w_1 and w_2.

MAIN: The logic of this state machine is similar to the main state machine in the `DNS` redirection use case. We sequentially go from one state to another. The complexity is hidden in the nested state machines that we execute using the `call` primitive.
The previous use case was totally symmetric, since the two nodes were sending the same request `r` to a `DNS` server. Here, the symmetry is broken as the two nodes act either as the client or the server. Nevertheless, we try to reduce the asymmetry of this test to only one variable, referred to as the **role**.
The role does not appear on the main state machine, so as to maintain a clean symmetry at the highest level. The nested state machines can retrieve the role that the master assigned with the help of the `get_file_parameter` primitive. This primitive can parse and retrieve the value of a specific variable from a configuration file transmitted by the master to launch a test campaign.

VARIABLES_SET_UP: This state is similar to the one with the same name in subsection 4.1. If the state machine has the role of the client, it creates a `TCP SYN` packet with the destination `IP` set as the server's `IP` address. The destination port is set to 80 for the `HTTP` protocol. The source port and the sequence number are random.
Independently of the role of the state machine, we define the name of the data channel queue to indicate at which place the received packets are stored. We also assign a filter to this queue so that it only listens for the `TCP` replies based on the source port. When this operation is done, the state machine starts to listen for incoming packets and stores the ones that are not filtered out. It forwards the name of the queue to the next state machine so that it can use it to process received packets.

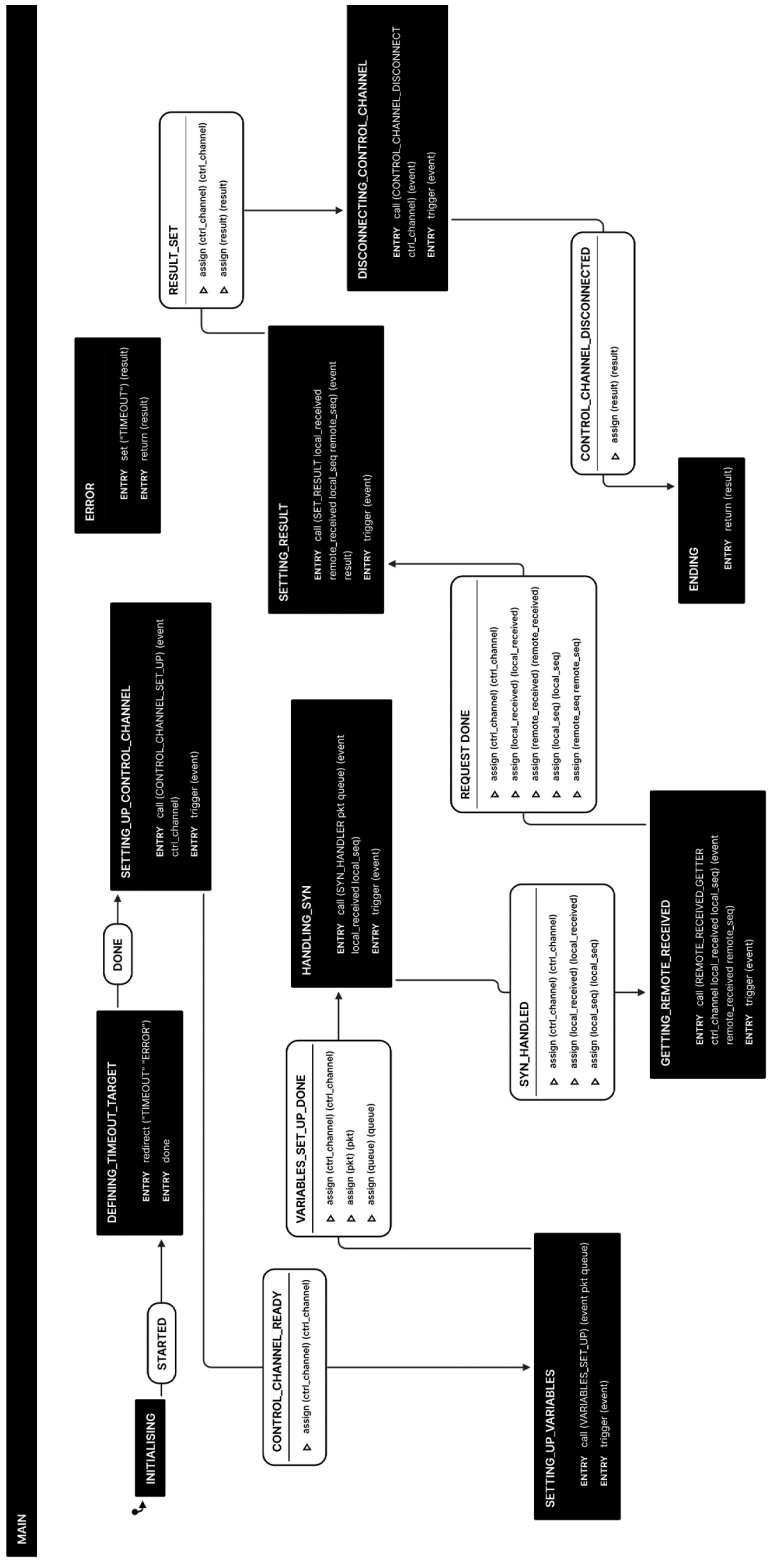


Figure 9. Main Finite State Machine in the case of a TCP transparent proxy.

SYN_HANDLER: This state sends a `SYN` packet from w_1 to W_2 and processes the first packet recorded using the name of the queue given as input parameter.
If the node that executes the program is the client, it sends the `SYN` packet to the remote worker with the `send` primitive and stores its sequence number into a *sequence* variable. This operation is only necessary for the client as the server does not need to send any packet.
Both the client and the server then start to process the packets received on the data channel queue. The program triggers a `packet_available` event as soon as a packet is available in the queue. They declare a new variable *received* and store a 1 into it to indicate that they received a reply. Since we want to compare the initial *sequence* number of the `TCP` session for both nodes, we also extract and store the sequence number of the `SYN` packet into the sequence variable for the server.
Conversely, if the node does not receive any packet, a `timeout` event is raised and we store a 0 into the received variable to indicate that the local state machine did not receive any reply. If it is the case of the server, the sequence variable holds the value `null`.
The nested state machine returns the received and *sequence* variables to the main state machine that then forwards them to the next state.

REMOTE_RECEIVED_GETTER: This state is responsible for the exchange of data between the two worker nodes.
It calls a nested state machine that sends one **sync message** to first synchronise with the remote worker. Inside this message, it stores the received and sequence local variables. The state machine then waits for the remote variables contained within a sync message in the control channel queue, using the `wait_sync_signal` primitive. As soon as a sync message is available on the queue, the state machine triggers a `sync_available` event to resume program execution. The nested state machine extracts the variables stored in the sync message and returns them to the main state machine. As a result the program now knows everything about what happened locally and remotely when the client sent a `SYN` packet.

RESULT_SETTER: The final step of the pipeline before disconnecting the control channel. The two machines use the local and remote values, along with their respective role in the scenario to infer the presence of a proxy and its category.
The nested state machine called for this task only relies on the **guard** mechanism within the transitions to provide the final result. If the server node set its local received variable as 0, then the state machine concludes that there is a sequential proxy. Otherwise, it checks if the client node set its local *received* variable as 1. In that case, the program infers that an asynchronous proxy is between the two nodes. Finally, if the local and remote *received* variables

are different, we have a synchronous proxy; when they are equal the program says that no proxy is detected.
This represents the final result that we can combine with other test cases afterwards to create a more fine-grained *test campaign*.

SQUID: In order to show how we can apply such a use case, we use our solution to find to which category a Squid transparent proxy belongs to. Fig. 10 represents the architecture for this concrete example. w_1 takes the role of the client that wants to categorize s, the Squid transparent proxy that acts as a man-in-the-middle between w_1 and the Internet. W_2, on the other hand, takes the role of the server. m is the master node that assigns the test case presented across this use case and P is the proxy used to initialise the control channel between w_1 and w_2. The firewall redirects TCP packets on port 80 to W_2 and blocks all other incoming traffic. The test case for this example determines that a **sequential** proxy is present. To verify that this result is correct, we use *Wireshark* to analyse the packets sent between the two TCP connections. The observations validate the conclusion given by our solution. Indeed, we do not observe any packet between s and W_2. On the other hand, we can observe many SYN−ACK retransmissions that go to w_1 with the IP address of W_2, but which are in fact generated by s rather than W_2. This phenomenon is due to the fact that w_1 does not acknowledge these SYN−ACK packets.

This example illustrates how we could use the state machine to classify a Squid transparent web proxy. We are in the process of reproducing this experiment on many different proxies with a test campaign specifically implemented for this purpose. For the sake of brevity, we defer the presentation of the results for some further workfuture study.

By means of these two use cases, we have demonstrated how we addressed the following specifications, as presented in subsection 3.2:

1. Ease of use
2. Ease of deployment
3. Modularity
4. Scalability
5. Flexibility

In these test cases, we have ensured ease of use, since an end user does not require any knowledge concerning how DNS and TCP packets work. The user only had to contact the master node M with the name of the test they wished to perform. The architectures we presented are easy to deploy and distribute, as we only need the master node to add and remove nodes (proxies or workers) and be aware of their status. For instance, we described how these use cases work with two worker nodes and a proxy P. They did not need any interaction with M after being assigned the test to run. The use cases highlight how our solution is modular and scalable. We only need the primitives present within our implementation to run a large variety of tests. Furthermore, we leveraged nested state machines to reuse two entire state machines. Finally, we showed how we ensure flexibility

as – in the case of the TCP proxy scenario – the data channel is between the two worker nodes. This is not the case when we want to verify whether a semi-active component redirects our DNS requests.

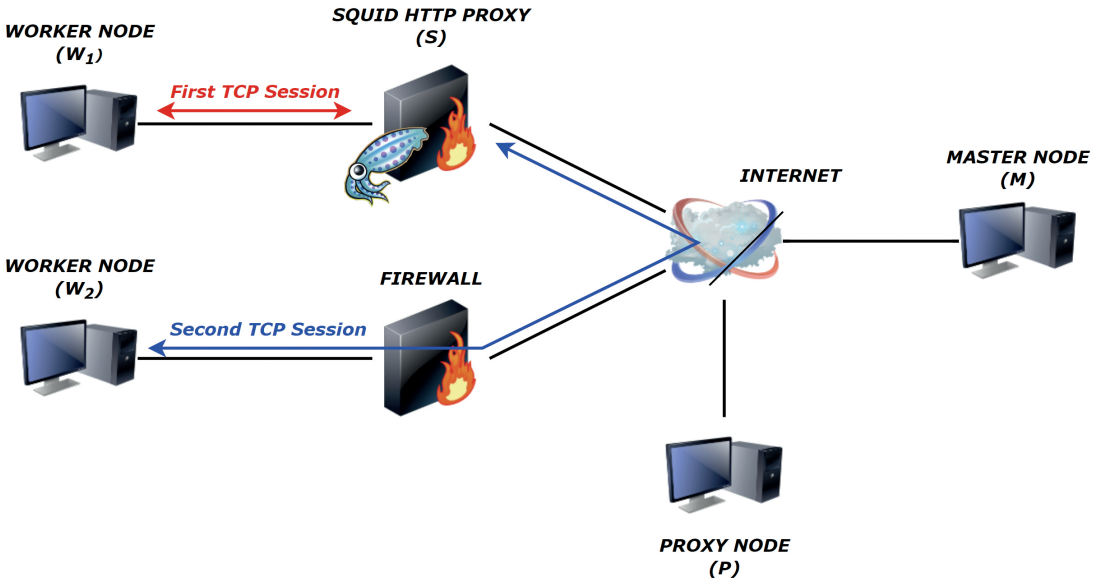


Figure 10. Architecture involving a Squid transparent TCP proxy.

5. Conclusion and future research

In this work, we have presented the foundation of a collaborative platform whose focus is to perform discovery and conformance verification of semi-active elements in network communications. In the future, we wish to make this platform open to anyone, available all over the world and with a large range of protocols supported. For this purpose, we are currently in the process of creating new scenarios, including (but not limited to) WAF discovery, firewall conformance verification, man-in-the-middle attack detection, fuzzing or HTTP smuggling execution, etc. Thus, we hope to have aroused the interest of the readers and welcome the contributions of anyone willing to lend us a hand. We particularly seek for collaborations in the development of the new modules that would make our framework more powerful, and also in the deployment of our platform around the world. The end goal we have in mind is to provide a worldwide map of the semi-active elements that may interfere with our daily communications.

Funding

This work was supported by King Abdullah University of Science and Technology (KAUST): Thuwal, Saudi Arabia.

References

[1] C. Kreibich, N. Weaver, B. Nechaev, V. Paxson, "Netalyzr: Illuminating the edge network," Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, 2010, pp. 246–259, doi: 10.1145/1879141.1879173.

[2] N. Weaver, C. Kreibich, M. Dam, V. Paxson, "Here be web proxies," in *Passive and Active Measurement. PAM 2014. Lecture Notes in Computer Science*, M. Faloutsos, A. Kuzmanovic, Eds. Cham: Springer, 2014, pp. 183–192, doi: 10.1007/978-3-319-04918-2_18.

[3] A. Vitale, M. Dacier, "Inmap-t: Leveraging TTCN-3 to test the security impact of intra network elements," *Journal of Computer and Communications*, vol. 9, pp. 174–190, 2021, doi: 10.4236/jcc.2021.96010.

[4] G.F. Lyon, *Nmap network scanning: The official Nmap project guide to network discovery and security scanning*. Sunnyvale, CA: Insecure. Com LLC, 2008.

[5] P. Baran, "On distributed communications networks," *IEEE Transactions on Communications Systems*, vol. 12, no. 1, pp. 1–9, 1964, doi: 10.1109/TCOM.1964.1088883.

[6] D.W. Davies, K.A. Bartlett, R. A. Scantlebury, P.T. Wilkinson, "A digital communication network for computers giving rapid response at remote terminals," Proceedings of the first ACM Symposium on Operating System Principles, 1967, pp. 2.1-2.17, doi: 10.1145/800001.811669.

[7] J.H. Saltzer, D.P. Reed, D.D. Clark, "End-to-end arguments in system design," *ACM Transactions on Computer Systems (TOCS)*, vol. 2, no. 4, pp. 277–288, 1984, doi: 10.1145/357401.357402.

[8] R. Oppliger, *SSL and TLS: Theory and practice*, 2nd ed. Norwood, MA: Artech House, Inc., 2016.

[9] S. Frankel, S. Krishnan, "IP security (IPSec) and Internet key exchange (IKE) document roadmap," RFC, vol. 6071, pp. 1–63, 2011, doi: 10.17487/RFC6071.

[10] F. Callegati, W. Cerroni, M. Ramilli, "Man-in-the-middle attack to the https protocol," *IEEE Security & Privacy*, vol. 7, no. 1, pp. 78–81, 2009, doi: 10.1109/MSP.2009.12.

[11] B. Aziz, G. Hamilton, "Detecting man-in-the-middle attacks by precise timing," in 2009 Third International Conference on Emerging Security Information, Systems and Technologies, 2009, pp. 81–86, doi: 10.1109/SECURWARE.2009.20.

[12] Y. Mirsky, N. Kalbo, Y. Elovici, A. Shabtai, "Vesper: Using echo analysis to detect man-in-the-middle attacks in LANs," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1638–1653, 2018, doi: 10.1109/TIFS.2018.2883177.

[13] M. Usama, M. Asim, S. Latif, J. Qadir, A. Al-Fuqaha, "Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems," in 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), 2019, pp. 78–83, doi: 10.1109/IWCMC.2019.8766353.

[14] Z. Trabelsi, K. Shuaib, "Nis04-4: Man-in-the-middle intrusion detection," *IEEE Globecom 2006*. pp. 1–6, 2006, doi: 10.1109/GLOCOM.2006.282.

[15] S. Miller, K. Curran, T. Lunney, "Traffic classification for the detection of anonymous web proxy routing," *International Journal for Information Security Research*, vol. 5, no. 1, pp. 538–545, 2015, doi: 10.20533/IJISR.2042.4639.2015.0061.

[16] M. Marlinspike. (2009). *New tricks for defeating SSL in practice.* [Online]. Available: https://www.blackhat.com/presentations/bh-dc-09/Marlinspike/BlackHat-DC-09-Marlinspike-Defeating-SSL.pdf. [Accessed: Sep. 28, 2022].

[17] E. Chiapponi, M. Dacier, O. Thonnard, M. Fangar, V. Rigal, "Badpass: Bots taking advantage of proxy as a service," in I*nformation Security Practice and Experience. ISPEC 2022*. Lecture Notes in Computer Science, C. Su, D. Gritzalis, V. Piuri, Eds. Cham: Springer, 2022, pp. 327–344.

[18] M. Champion, M. Dacier, E. Chiapponi, M. Fangar, V. Rigal. (2022). *Immune: Improved multilateration in noisy environments, Eurecom.* [Online]. Available: https://www.eurecom.fr/publication/7065https://www.eurecom.fr/publication/7065. [Accessed: Oct. 24, 2022].

[19] V. Paxson. (2022). Personal communication.

[20] I. Schieferdecker, A.G. Vouffo-Feudjio, “The testing and test control notation TTCN-3 and its use,” *Formal Methods for Industrial Critical Systems: A Survey of Applications*, 2012, pp. 205–233, doi: 10.1002/9781118459898.ch10.

[21] M. Roesch, “Snort: Lightweight intrusion detection for networks,” Proceedings of LISA '99: 13th Systems Administration Conference Seattle, Washington, 1999, pp. 229–238.

[22] S. Bansal, N. Bansal, “Scapy-a python tool for security testing,” J*ournal of Computer Science & Systems Biology*, vol. 8, no. 3, p. 140, 2015, doi: 10.4172/JCSB.1000182.

[23] Statelyai. (2021). Xstate. [Online]. Available: https://github.com/statelyai/xstate. [Accessed: Oct. 24, 2022].

[24] M. Myers, R. Ankney, A. Malpani, S. Galperin, C. Adams, “X.509 internet public key infrastructure online certificate status protocol – OCSP,” *RFC*, vol. 2560, pp. 1–23, 1999, doi: 10.17487/RFC2560.

[25] Ansible. (2022). *Ansible*. [Online]. Available: https://github.com/ansible/ansible. [Accessed: Oct. 24, 2022].

The (Il)legitimacy of Cybersecurity. An Application of Just Securitization Theory to Cybersecurity based on the Principle of Subsidiarity

Johannes Thumfart | **Research Group Law, Science, Technology and Society (LSTS),Department of Metajuridica, Faculty of Law and Criminology, Vrije Universiteit Brussels, Belgium; International security management at the Faculty of Police and Security, Berlin School of Economics and Law, Germany**, ORCID: 0000-0003-4337-2990

Abstract

The application of securitization theory to cybersecurity is useful since it subjects the emotive rhetoric of threat construction to critical scrutiny. Floyd's just securitization theory (JST) constitutes a mixture of securitization theory and just war theory. Unlike traditional securitization theory, it also addresses the normative question of when securitization is legitimate. In this contribution, I critically apply Floyd's JST to cybersecurity and develop my own version of JST based on subsidiarity. Floyd's JST follows a minimalistic and subsidiary approach by emphasizing that securitization is only legitimate if it has a reasonable chance of success in averting threats to the satisfaction of basic human needs. From this restrictive perspective, cyber-securitization is only legitimate if it serves to protect critical infrastructure. Whilst Floyd's JST focuses exclusively on permissibility and needs instead of rights, I argue that there are cases in which states' compliance with human rights obligations requires the guarantee of cybersecurity, most importantly regarding the human right to privacy. My version of JST is also based on the principle of subsidiarity, in the sense that securitization should always include stakeholders directly affected by a threat. To strengthen this kind of subsidiarity, focused on the private sector, I argue for the legitimacy of private active self-defence in cyberspace and emphasize the importance of a 'whole-of-society approach' involving digital literacy and everyday

Corresponding author:
Johannes Thumfart, Research Group Law, Science, Technology and Society (LSTS), Department of Metajuridica, Faculty of Law and Criminology, Vrije Universiteit Brussels, Belgium; International security management at the Faculty of Police and Security, Berlin School of Economics and Law, Germany; ORCID: 0000-0003-4337-2990; johannes.thumfart@vub.be

security practices. Moreover, I argue that far-reaching securitization on the nation-state-level should be avoided, particularly the hyper-securitization of the digital public sphere, following unclear notions of 'digital sovereignty'.

Keywords
cybersecurity dilemma, desecuritization, digital sovereignty, securitization, securitization theory, societal security dilemma

Cite this article as: J. Thumfart, "The (Il)legitimacy of Cybersecurity. An Application of Just Securitization Theory to Cybersecurity based on the Principle of Subsidiarity," ACIG, vol. 1, no. 1, pp. 119–147, 2022, DOI: 10.5604/01.3001.0016.1093

1. Introduction

Cybersecurity is a particularly contested branch of security, since it relates to a sociotechnical environment that is tightly interwoven with digital civil society and the global free flow of information and services. Correspondingly, cyberspace was historically linked to a cyber-libertarian political culture [1, 2]. From this perspective, it is hardly surprising that the issue of cybersecurity is increasingly discussed within the critical framework of securitization theory [3–6].

Securitization theory was developed by proponents of the Copenhagen School in the 1990s [7]; this theory originally examined 'securitizing speech acts' by which political leaders identify or construct a threat to a 'referent object of security' and promote 'exceptional measures' in order to avert or prevent this threat. This approach relies on the philosophical concept of speech acts that describes utterances which have a high social impact, for example, oaths, declarations of war, and calls to arms [8]. Rather than constituting a neutral methodological turn, the Copenhagen School's shift toward the examination of the role of language in the social construction of security is inherently critical since it is decidedly anti-positivist: its primarily linguistic analyses denaturalize the emotive rhetoric, subconscious fight-or-flight responses, and populist impulses connected with security issues [9, 10].

It is debatable whether securitization theory, which is primarily connected with international conflict and which emphasizes 'extraordinary measures' (including severe forms of coercion), can be applied to cybersecurity at all, which, in turn, is usually connected with everyday problems and focused on civilian technological routine [3]. Regardless of these difficulties, which will be addressed in detail, there are significant benefits from applying it in this manner. Cybersecurity is often related to various degrees of threat inflation, driven by widespread fears regarding rapid technological development on the one hand, and the concrete interests of security experts in the public and private sectors in increasing their wealth and/or power on the other hand, which has been described as the "cyber-industrial complex" [11] or "military-digital complex" [12]. This process has been criticized as "hyper-securitization", in the sense of exaggerated securitization, [3, 6], involving "inflationary and sensationalist danger hyping" [3] and "the rise of militaristic rhetoric around digital threats" [13].

Alongside the misallocation of resources due to such threat inflation, securitization can have undesirable effects on two different levels. First, there is the danger of a 'cybersecurity dilemma': when engaging in cyber-securitization in the sense of creating deterrence by threatening to defend forward, states increase their own security to the detriment of others and thereby destabilize the international system [14]. A similar dilemma arises when states promote their security by submitting civil society to widespread digital surveillance [15]. Second, particularly since cybersecurity is closely related to the digital public sphere, it can lead to the securitization of the digital public sphere [16], which conflicts with the democratic core value of freedom of speech [17]. In consequence, the securitization of cyberspace produces a version of the 'societal security dilemma' [4, 18]: Framing cyber interference as an international security issue shifts the focus away from resolving the domestic social tensions that create vulnerabilities to cyber interference in the first place.

Because these problems require a normative approach toward securitization, they constitute a suitable application of Rita Floyd's Just Securitization Theory (JST) [19–22], a critical offshoot of the Copenhagen School's approach. Floyd criticizes the Copenhagen School approach as "analytically strong" but "normatively weak" [22]. Her own JST is not exclusively focused on social constructivism but has a more empirical dimension and is normatively productive, attempting to answer the normative question of the legitimacy of securitization. Applying JST to cybersecurity will, therefore, allow a distinction to be drawn between legitimate and illegitimate forms of securitization in cyberspace, with a special focus on the avoidance of hyper-securitization, the cyber security dilemma, and the societal security dilemma. This also includes a normative assessment of different approaches to digital sovereignty [23].

Floyd's JST follows a generally minimalistic and subsidiary approach by emphasizing that the legitimacy of securitization is only given if it stands "a reasonable chance of success" in averting threats to the satisfaction of "basic human needs" [19]. Likewise, she poses the complementary questions of when desecuritization is morally obligatory and how desecuritization should be implemented; 'desecuritization' implies the reversal of securitization, "the shifting of issues out of emergency mode and into the normal bargaining processes of the political sphere" [7].

Table 1. Main differences between Floyd's JST and my JST

	JUST INITIATION	JUST CONDUCT	JUST TERMINATION
FLOYD'S JST	• Focused on basic human needs instead of rights • Focused on permissibility	• Focused on physical threats • Focused on state actors • Contains elements of retributive justice • Considers intentions behind threats	• Securitization and desecuritization are understood as a dichotomy • Focused on sustainable securitization
MY JST TAILORED TO CYBERSECURITY	• Focused on protecting individual rights • Including states' obligation to guarantee security • Emphasizes the knowledge of stakeholders directly affected by a threat	• Considers non-physical threats • Emphasizes stakeholders' everyday security practices • Includes private actors' right to active self-defence • Separates retributive justice from securitization because it requires judicial procedures • Intentions behind threats are irrelevant • No financial or political gains should be connected with securitization	• Securitization and desecuritization are understood as continuous • Hyper-securitization is avoided, and sustainability is guaranteed by subsidiarity

During the discussion of the applicability of Floyd's JST to cybersecurity, I develop my own version of JST (Tab. 1.). First, I extend Floyd's focus, which is restricted to permissibility regarding securitization, to also include nation states' moral and legal obligation to securitize in order to protect human rights relevant to cyberspace, above all the right to privacy. Furthermore, I include private actors and everyday security practices [24] to a greater degree than in Floyd's original JST. Following Hansen and Nissenbaum's definition [6], everyday security practices include the practical knowledge of direct stakeholders and the consideration of the relevance of human behaviour to cybersecurity issues, for example, regarding passwords or phishing mails. The focus on such practices enhances the principle of subsidiarity that is already pre-figured in Floyd's original account. I argue that hyper-securitization can best be avoided by involving the stakeholders most directly affected by a threat in the decision-making processes, and by ensuring that securitization can be enacted autonomously by these stakeholders, provided that they have the necessary legal and technical competencies. In this context, I also emphasize the importance of a right to active self-defence in cyberspace [25]. In general, I argue that the Copenhagen School's dichotomous understanding of securitization and

desecuritization, which is also reflected in Floyd's JST, is not applicable to the more continuous dynamics of securitization and desecuritization in cyberspace.

In the second section, I provide a critical discussion of the literature focused on the question of how securitization theory, which stems from the field of international conflict, can be applied to cybersecurity at all, which also involves international conflicts but is mostly focused on technological routine in everyday situations. The subsequent sections follow the structure of JST, which is tripartite in terms of just initiation, just conduct, and just termination. This tripartite nature corresponds to the partition of traditional just war theory in *jus ad bellum* (the right to war), *jus in bello* (rights in war), and j*us post bellum* (rights after war). In the third section, I raise the question of 'just initiation' regarding cybersecurity, which is largely determined by the question of which kind of threat allows for securitization or even morally requires it; in the fourth section, I pose the question of the 'just conduct' in cybersecurity, which involves the development of normative criteria regarding the concrete measures taken during securitization processes; in the fifth section, I pose the third and most difficult question, namely that of desecuritization. These sections are loosely divided into thematic subsections. The whole discussion is followed by a conclusion.

2. Literature Review focusing on the Application of Securitization Theory to Cyberspace and the Principle of Subsidiarity

2.1. The Incompatibilities of Securitization Theory and Cybersecurity

Since JST is a comparably new concept in Security Studies, until now it has only been applied in singular cases [26]. Besides a brief mention of Floyd's research in an article related to cybersecurity [27], JST has not yet been applied to cybersecurity at all. This is the most obvious research gap that this contribution addresses. As will be discussed in detail in Subsection 3.1, applying JST to cybersecurity has the advantage of considering the critical and constructivist approach of the Copenhagen School, whilst adding a normative and productive element to it.

In contrast to its recent offshoot, JST, securitization theory has been applied to cybersecurity frequently, e.g., [3–6]. The problems with the application of securitization theory to cybersecurity have been most extensively illuminated in Dunn Cavelty's discussion of the research literature [3]. Securitization theory originally developed from a focus on international conflict and understands security primarily as involving the use of force or the threat of force; according to Dunn Cavelty, the theory's genealogy from international conflict determines its focus on exceptional measures, i.e., measures outside of the 'normal' political order of liberal states, including trade-offs between fundamental rights or even the use of force. In contrast to this, cybersecurity is more focused on technological routines, which constitute the everyday situation rather than exceptional measures. She writes:

When focusing on security that is no longer primarily about threats and battles against an enemy, but is characterized by an inward-looking narrative about vulnerabilities, it becomes necessary to question the perception of security as 'exceptional' and linked to 'extraordinary' means [3].

This observation represents an accurate caveat to this paper's basic attempt to apply JST to the issue of cybersecurity. After all, JST is an offshoot of the Copenhagen School approach and shares its origin in international relations theory. In turn, cybersecurity can involve conflicts between states, but it usually does not, particularly not in the sense of a direct clash between state actors. Non-state actors play an important role in international cybersecurity, as proxies or mercenaries (such as state-sponsored hacker groups) or as agents acting with various degree of independence (for instance, non-state hacker groups) [28].

But even if relevant cases involve a direct confrontation between states, they usually exhibit low intensity and do not pass the critical threshold of the use of force [29–32]. A particularly well-known example of this kind of conflict is the Russian meddling with the digital public spheres of Western countries, which barely exceeds the spread of propaganda or cyber espionage and will be discussed in detail in section 4. Other phenomena such as cybercrime largely involving private actors only have a small significance to international security in the narrow sense.

2.2. The Compatibilities of Securitization Theory and Cybersecurity

Whilst the incompatibilities of the cybersecurity discourse and the Copenhagen School approach mentioned in the previous subsection are important, one should not overemphasize the degree to which classical securitization theory is exclusively applicable to international conflicts. First, regarding its context, securitization theory stems from the post-Cold War period in the 1990s. This era is characterized by a shift in the security discourse from a relatively simple paradigm of state-centric bipolarity to other, more complex problems, such as humanitarian interventions and terrorism.

When Buzan and Wæver developed securitization theory, they did so with the explicit intention of constructing a pathway between Habermas's discourse-oriented political theory and Schmitt's authoritarian theory of the state, focusing on the state's ability to implement exceptional, even unconstitutional measures in emergency situations [7], [33]. Habermas' discursive legitimatization of democracy does not relate to international conflict [34]; neither does the part of Schmitt's theory that includes his endorsement of the 'state of exception', which rather addresses the 'inner enemy' [35]. In liberal states, this inner enemy could be terrorists. Of course, terrorism is relevant to international security, but its domestic, societal, and cultural aspects demonstrate that securitization theory is familiar with issues that go beyond a paradigm focused on conflicts between states and involves many 'softer' and more complex mechanisms of creating societal security, rather than exceptional measures in the traditional sense [36]. Furthermore, securitization theory has been applied to ethnic conflict [18],

HIV/AIDS [37], and human and drug trafficking [38], which all lie beyond the traditional scope of international conflict.

Another aspect of Dunn Cavelty's analysis of the problems with applying classical securitization theory to cybersecurity is the fact that the former is primarily focused on securitizing speech acts, i.e., the rhetorical process of declaring an issue relevant to security. This would imply that securitization theory can hardly be applied to cybersecurity, which is focused on technological procedures rather than on mere rhetoric, and is often not even publicly discussed, a trait that Dunn Cavelty characterizes as "non-discursive practices" [3].

Whilst this terminology is somewhat confusing, since it seems to suggest that there is practice without discourse, it probably has to be understood in the sense that cybersecurity is often exclusively the domain of "technical experts, rather than other political actors" and, therefore, not necessarily publicly discussed [3]. In underlining the non-discursive features of cybersecurity in this sense, Dunn Cavelty follows Hansen and Nissenbaum's seminal critique of "technification that depoliticizes" [6].

However, the Copenhagen School's emphasis on language also needs to be relativized regarding its methodological implications. From a methodological perspective, it must be underlined that traditional securitization theory focuses on discourses and speech acts, which both go beyond language itself. Traditional securitization theory cited Foucault and Austin as the theorists who coined these concepts [7]. Although the early work of the Copenhagen School indeed focused mostly on language [39], the original Foucauldian understanding of discourses includes theoretical and practical features [40]. Likewise, according to Austin, speech acts are defined by their relation to extra-linguistic practices and contexts [8]. The threshold between theory and practice is particularly permeable in the case of cybersecurity, which involves programming languages, algorithms, and workflows that are, by their very nature, situated on the border between theory and practice and can be understood in terms of Austin's speech acts [41] or in terms of Foucault's discourses [42]. Hence, cybersecurity practices fall within the scope of traditional securitization theory, if *discourses* and *speech acts* are understood acknowledging the full range of these concepts.

In a recent collaborative paper with Egloff, Dunn Cavelty cited the Swiss model of subsidiarity as a possible means of bridging the conceptual conflict between cybersecurity as everyday security practices (often enacted by private stakeholders and neither including 'extraordinary measures' nor public speech acts) and state-level security agendas that tend to 'hyper-securitize', involving public rhetoric [24]. According to this principle, "a central authority should perform only those tasks which cannot be performed effectively at a more immediate or local level" [24]. Following Hansen and Nissenbaum's seminal definition, everyday security practices have to be understood in a double sense: first, they apply the practical knowledge of direct stakeholders to securitization processes; second, they consider the fact that human negligence, for example, regarding passwords or phishing mails, ranks among the most important security threats [6]. In my own account of JST, I will develop this notion of subsidiarity regarding cyber-securitization.

In summary, I argue that the continuities between securitization theory and cybersecurity are more important than their obvious incompatibilities. These continuities are: the definition and speculative construction of threats to a valued object of security, the discourses about and practical production of adequate means to avert these threats (including the attribution of resources), and the corresponding highly emotional (and thus dangerous) fight-or-flight responses. From this perspective, the routine aspects of cybersecurity and the exceptional aspects of cybersecurity can be understood within a securitization framework.

3. What is Just Initiation regarding Cybersecurity?

3.1. JST as a remedy against the Copenhagen School's Normative Weakness

Traditional securitization theory is largely critical of securitization since it focuses on putting the naturalization of securitization into question. The classical formula of the Copenhagen School defines the securitizing speech act as a discursive operation by which a securitizing actor (usually the government) justifies exceptional measures to avert an existential threat from a valued referent object of security (e.g., critical infrastructure, sensitive information, the nation, human livelihoods) in front of an audience (usually the public but also expert circles) [7]. Subsequently, the Copenhagen School analysed these securitizing speech acts (also called 'securitization moves') primarily with a focus on the rhetoric of political elites – but, as argued above, the school's methodological focus on discourses and speech acts allows, in principle, for far more practical applications.

The linguistic focus of the Copenhagen School produces a significant degree of moral relativism and incompatibility with the needs of practitioners to gain normative orientation: on the one hand, its methodological constructivism implicitly denies the possibility of security threats that simply "exist 'out there'" [19]; on the other hand, it might be misunderstood as reducing the often brutal measures of securitizing actors to mere rhetorical operations. In Floyd's words: "While analytically strong, the Copenhagen School's theory is normatively weak" [22].

Floyd tackles these weaknesses of the Copenhagen School approach, its moral relativism and lack of practical applicability, by two adjustments to classical securitization theory: first, she focuses on 'security actions' instead of 'securitizing speech acts' or 'securitizing moves', i.e., not merely the rhetoric but actual measures undertaken in acts of securitization. She writes: "Securitization is possible without the securitizing move but not without security action" [43]. Again, comparable to Dunn Cavelty's idea of "*non-discursive practices*" discussed in subsection 2.2, this must raise eyebrows but is probably meant in a sense that securitization measures are not necessarily publicly addressed or discussed. Whilst it is debatable whether Floyd is accurate to assume that exclusively rhetorical securitization cannot constitute securitization, concrete securitization measures (involving trade-offs between fundamental rights or even the use of force) are obviously to be discussed much more critically than mere rhetoric.

Furthermore, Floyd argues that, due to its de-naturalizing approach focused on social constructivism, the Copenhagen School has, strictly speaking, no concept of objective threats. Moreover, if securitization theory does not distinguish between objective threats and their rhetorical construction, then all securitization moves have the same degree of legitimacy or the same lack of legitimacy. From such a dangerously relativist perspective, the US's attack on Iraq in order to avert the imaginary threat of Weapons of Mass Destruction has the same value as Sweden and Norway joining NATO to avert the real threat of Russian aggression; or, to take an example from cybersecurity, the hysterical securitization due to the unfounded fear of Y2K [4] would have the same value as the necessary securitization regarding WannaCry, which was likely to cause physical damage by attacking hospitals [44]. Floyd writes:

> An exclusive focus on the constructedness of security means (...) that securitization scholars tend to ignore whether or not the threats that inform securitization are real or otherwise. (...) A better strategy is to begin by (...) judging the objective existence of a threat, because unless there is a real threat, securitization is most definitely the wrong political and ethical choice [19].

Regarding the question of which objective threats to which referent objects make securitization legitimate, Floyd pursues a restrictive and subsidiary approach. According to her, the legitimacy of securitization is only given if it has "a reasonable chance of success" in averting threats to the satisfaction of "basic human needs" [19]. This minimalism is highly ambivalent if applied to the usually non-physical realm of cybersecurity. On the one hand, it is certainly possible to justify the securitization of critical infrastructure from this perspective, e.g., networks related to water, energy, and food supply. On the other hand, this focus on existential human needs may constitute an overly restrictive threshold, considering the non-physical scope of the vast majority of problems related to cybersecurity and the largely private sector-oriented and more quotidian, routine-driven, and civilian nature of cybersecurity, which does not necessarily require such a high threshold as it does not necessarily involve trade-offs between fundamental rights.

But it is worth taking a closer look at Floyd's argument. The advantage of her restrictive focus on concrete human needs as the only legitimate referent objects of security becomes particularly evident if contrasted with other possible referent objects, most importantly, the state. Unlike human beings, the state does not have a moral quality *a priori* because a state might be a dysfunctional dictatorship committing crimes against humanity. Additionally, particularly from a human rights-centric perspective based on the rule of law, the legitimacy of the state depends on whether it complies with human rights [43, 45].

From the perspective of Floyd's JST, which is not focused on rights but on output legitimacy regarding the satisfaction of basic human needs, this allows the following assessment: whilst Floyd's JST includes the possibility that nondemocratic regimes which guarantee the satisfaction of the basic needs of their citizens can be legitimate referent objects of

securitization, this is not the case regarding democratic regimes that cannot guarantee this. One might add that Floyd's JST is too focused on output legitimacy in this regard. Emphasizing human needs instead of human rights might open the door for utilitarian reasoning, trading legitimacy for efficiency. However, Floyd argues that speaking of human needs opens the possibility of including referent objects in JST that do not involve actors and do not have a legislative function, most importantly the ecosystem, which will be discussed in relation to cybersecurity in subsection 3.3 [43].

3.2. JST as a Human-Centric approach to Cybersecurity

If applied to cybersecurity, Floyd's focus on human beings is particularly advantageous in the sense that it constitutes a principle of subsidiarity. It constitutes a "human-centric approach to cybersecurity", as defined by Deibert, to actively counteract a fixation on national security [46]. Frequently, the nebulous and largely rhetorical discourses about 'digital sovereignty' [23] and national cyber security [47] promote 'hyper-securitization' without providing a clear referent object of security. Furthermore, claims to 'digital sovereignty' inherently create what Mueller calls the 'cyberspace jurisdiction paradox' [48]: in the 'post-territorial' environment [46] of digital networks, claims to exercise territorial jurisdiction necessarily transcend borders and have extraterritorial features, such as exemplified by the digital aspects of the 'Brussels Effect' [49] and the 'Beijing Effect' [50]. This means that both the EU and China are exercising forms of extraterritorial jurisdiction in and through cyberspace: the EU in a regulatory sense, China by standard setting but also via globally available Chinese apps such as Alipay, WeChat Pay and TikTok, which share data with the Chinese government [51].

Analogue to these practices of extraterritorial jurisdiction and extraterritorial *de facto* control, the escalating rhetoric about 'digital sovereignty' [23] creates a 'cybersecurity dilemma' of states striving for an enhancement of their own security by threatening to defend forward, begetting international conflicts. Take for example NATO's doctrine that cyber-attacks can be interpreted as triggering a collective response according to Article 5 of the charter, including kinetic responses [52, 53]. Whilst such doctrines might have a deterrent effect, they can obviously cause destabilization, particularly since attribution is notoriously contested in cyberspace [54]. Moreover, as Dunn Cavelty argues, hyper-securitization not only creates a cybersecurity dilemma on the level of international relations but also regarding the relationship between states and individuals, whose rights are often seriously affected by states' hyper-securitization, such as with mass surveillance [15].

From the perspective of Floyd's JST, discourses regarding national cybersecurity need to be critically examined with respect to whether security claims related to 'sovereignty' actually refer to existential human needs. After all, the notions of 'cyber sovereignty' and 'information sovereignty' emerged in the context of Chinese authoritarianism in the late 1990s [55]. If such claims are not related to the everyday reality of concrete human needs, then their legitimacy seems doubtful. This is particularly the case if

they involve restrictions on fundamental rights such as the right to privacy or the right to freedom of information, which includes the right "to receive and impart information and ideas of all kinds, regardless of borders" [56].

In turn, particularly because sovereign states are the ultimate guarantors of human rights, it is possible to conceive of forms of cyber-securitization following the paradigm of sovereignty that aim precisely at protecting these human rights. This is, for example, the case when claims to sovereignty are made to "draw a line" and protect individuals on a state's territory from becoming victims of digital transnational repression, by which authoritarian states reach into the territory of liberal states [57]. This will be discussed further in subsection 3.4.

3.3. JST, Posthuman Security, and the Participation of Civil Society

Floyd's understanding of the ecosystem as a legitimate referent object of securitization, by virtue of its functioning as a guarantee of the satisfaction of existential human needs, can also be applied to cybersecurity. Her approach in this regard can be understood in terms of 'posthuman security'. As has been argued by Mitchell in her essay on this issue [58], such an approach could be relevant to cybersecurity, which involves the protection of networks by focusing on their networked character. This means that complex digital networks require risk awareness in their own right because, by virtue of their mere complexity and high degree of connectivity, they increase the probability of improbable but dangerous 'black swan' events [59]. Moreover, these events have the potential to affect the whole world, starting with automatized financial markets, for instance, as was the case in the 'flash crash' of 2010 [60]. Hansen and Nissenbaum scrutinized such digital disasters from a critical perspective, yet they did not entirely dismiss their plausibility [6].

In particular, such 'black swan' events, which are impossible to predict, raise complex epistemological questions. Floyd misses the opportunity to apply her human-centric approach to these epistemological problems, which constitutes another application of the principle of subsidiarity in my version of JST. Rather than being merely determined by governmental experts, securitization processes should include the participation of relevant stakeholders to achieve maximum epistemic certainty. The participation of relevant stakeholders is particularly important regarding cyber-securitization. In the cybersecurity sector, states do not necessarily have the upper hand in regard to skills and expertise [61], [62]. In many cases, private companies and civil society actors are better informed about weaknesses, exploits, and possible ways of counteracting them.

Enabling the participation of a wide array of these stakeholders could constitute a powerful aspect of building resilience. For instance, the German Government regularly consults with Europe's largest association of hackers, the Chaos Computer Club [63]. The Swiss government is likewise pursuing an approach focusing on everyday security practices [24]. The EU is pursuing a "whole-of-society approach" [64]. Such bottom-up approaches could also be relevant to international cybersecurity, constructing

a private-public network of "distributed cyber deterrence" [25]. The US's Joint Cyber Defence Collaborative involving public-private partnerships is heading in this direction too [65]. If they include all levels of society, not merely an "invisible handshake" between Big Tech and governments [66], such participatory approaches could serve as a corrective force in securitization processes, making sure that fundamental rights and the interests of civil society are acknowledged by nation states, which are, chiefly due to their monopoly on violence, still the most powerful and most important securitization actors.

3.4. States' Obligation to Securitize

Particularly if securitization is connected to civil society, existential human needs, and fundamental rights on the legitimatory and epistemological level, in the way outlined in the previous subsections, it seems incoherent that Floyd's original account of JST focuses exclusively on permissibility, i.e., the question of when securitization *is allowed* [19]. If securitization is constituted in discourses which include the participation of relevant stakeholders, guarantee the satisfaction of existential human needs, and foster the enjoyment of human rights, then the emphasis on nation states' obligation to securitize not only represents a moral imperative but also a legal requirement. In a separate text addressing states' obligation to securitize, Floyd emphasizes different understandings of 'last resort' regarding exceptional measures in emergency situations that could bring about such an obligation [22]. But Floyd discusses this issue exclusively from an effects-based perspective and from a perspective that emphasizes the non-quotidian nature of securitization (the latter will be critically discussed in section 5.)

However, based on a more quotidian understanding of security and rights-based approaches, states are obliged to guarantee human rights on their territory, and this obligation clearly extends to cyberspace also. For instance, acts of digital transnational repression, in which authoritarian governments target exiled dissidents abroad, demand security measures since these actions threaten human rights on the host states' territory [67]. And such measures can most effectively be legitimized and communicated using the language of sovereignty, which is, in this sense, a tool for compliance with international obligations [57]. The EU's NIS Directive includes far-reaching obligations of member states to guarantee cybersecurity and to collaborate in this field [68]. Multilateral campaigns, such as the discontinued 'Clean Network Initiative', can be expected to increasingly include contractual obligations in regard to cybersecurity, in order to guarantee the free flow of information within international networks of trusted actors [69]. This constitutes the cybersecurity equivalent of the general economic trend toward so-called 'friend-shoring' [70].

These developing obligations of states in regard to cybersecurity also affect the private sector in an intermediate way, i.e., if legislation requires private actors to provide a certain degree of security, as is stipulated in the EU's NIS Directive. The relative proximity of public actors to the political ends of societies and the respective public debates on the one hand,

and the relative remoteness of private actors from such debates on the other hand, contribute to the illusion that cybersecurity by private experts is exempted from political discourse. Such technification and depoliticization [6] is enhanced by excluding private stakeholders from debates regarding the normative foundations of national cybersecurity. In turn, my version of JST, which is focused on subsidiarity, *re-politicizes* cyber-securitization by explicitly including the private sector and emphasizing its obligations within everyday security practices. Following Hansen and Nissenbaum's definition, everyday security practices have two dimensions, involving civil society as an "ambiguous partner and a potential threat": first, they utilize the practical knowledge of direct stakeholders in securitization; second, they consider the fact that human negligence, for example, regarding passwords or phishing mails, ranks among the most important security threats [6].

4. What is Just Conduct regarding Cybersecurity?

4.1. Cybersecurity, Physical Violence, and Intentionality

In the previous section, I have considerably extended the possibilities and responsibilities of states and private actors regarding cyber-securitization in comparison to Floyd's restrictive account of JST. This requires submitting the issue of 'just conduct' to critical scrutiny.

Following the orthodox definition of 'exceptional measures', Floyd argues that securitizing actors are entitled to suspend some human rights in securitization processes [43]. She argues against the legitimacy of similar competencies regarding cyber threats, since they usually do not involve physical violence. She writes: "when there is a direct lethal threat, securitization can involve lethal force." However, the removal of hackers "by means of lethal force on the part of the police or some special branch thereof, or even a military strike would be unjust, because they do not pose a direct threat to human life [43]."

Indeed, the cases in which hackers or cyber-attacks brought about direct physical harm are quite rare [71]. An attack involving flickering images used on epileptic victims was one rare incident in which hacking brought about direct physical violence [72]. Other exceptional instances, such as a ransomware attack on the British NHS, might have caused physical harm in a sense that is almost as direct [44]. Considering such exceptional cases that involve violence indirectly or more or less directly, Floyd's JST would certainly allow for securitization.

However, according to Floyd, this is not necessarily the case if the physically violent consequences of a cyber threat were unintended. She writes: "Agents who do not realize that their actions are (...) lethal to other people (are) (...) morally irresponsible for posing the unjust threat [43]." Whilst not entirely dismissing the possibility of legitimate securitization regarding such threats caused – but not intended – by an agent, Floyd emphasizes that such threats do not allow for establishing a "standard formula" of securitization. At first sight, this distinction between "agent-intended threats" and merely "agent-caused threats" seems to make sense. But introducing the category of intention constructs a nexus between morality

and securitization that is highly problematic. By discussing responsibility and intention, Floyd seems to have constructed JST as a punitive instrument related to retributive justice.

However, under the rule of law, punishment can only be enacted following legal procedures, which reach conclusions about how to judge the intentions of suspects that are crucial to determining the degree of guilt. Such proceduralism is particularly important because, due to their non-empirical nature, intentions are inherently difficult to judge. In contrast to this, securitization theory is primarily concerned with averting a threat *pre-emptively and immediately*, and this scope conflicts with such cumbersome proceduralism. Therefore, in the unlikely case that the actions of someone unintentionally pose a lethal threat, it seems to be legitimate to avert this threat by 'extraordinary measures' involving the degree of coercion necessary to avert the threat, regardless of the actor's intentions. From a security perspective focused on the aversion of an immediate threat and the protection of physical integrity, it simply does not matter whether an average person poses a lethal threat by accident, or a brilliantly strategizing terrorist poses a lethal threat intentionally. What changes is merely the fact that a one-time intervention is likely to be much more successful if a particular actor is intentionally posing a threat; and non-intended threats might be reoccurring despite the neutralization of one particular actor.

Moreover, according to Floyd's JST, the decisive criterion regarding the legitimacy of 'exceptional measures' involved in securitization is the degree to which it can be expected that a suspension of human rights (including severe coercion) has "a reasonable chance of success" in averting a threat [43]. In addition to this, she emphasizes proportionality, which is a typical feature of just war theories. This means that the rights violations caused by the degree of coercion used to avert the threat cannot be greater than the rights violations that can be reasonably expected to be caused by the threat. Furthermore, she writes that securitization should, in general, do "the least amount of overall harm possible" [43] Finally, she writes that securitization must be aimed at being reversed by desecuritization at one point and it needs to include measures to "avoid renewed (...) securitization" [43], which means that it should aim for sustainable stability.

4.2. Intelligence Operations and the Securitization of the Digital Public Sphere

Whilst the criteria discussed in the previous subsection are rather uncontroversial, it is far from clear how they would apply to cybersecurity, assuming that a cyber-attack usually does not involve physical violence. Cyber-attacks usually only threaten the right to physical and intellectual property, privacy, and freedom of speech. Thus, Floyd's own account of JST focusing on basic human needs is clearly only applicable here if one understands it in a non-physical manner, involving human rights.

However, Floyd makes an interesting argument that can be used to shed some light on the rationale behind this distinction and which is useful for the application of JST to cybersecurity. She argues that, unlike threats to property, physical threats can be legitimately subjected to securitization

processes because they inflict damage that cannot be restituted [43]. Whilst recognizing that Floyd's JST does not include the possibility of being applied to cybersecurity in relation to attacks without physically violent effects (regarding her own explicit assertions discussed in the previous subsection), it is worth asking how extensive a threat to technical infrastructure and complex digital networks would have to be to cause such irreversible damage. Similarly, Hansen and Nissenbaum emphasize the irreversibility of any damage inflicted as a feature that connects the protection of complex networks with the protection of the climate [6], which relates to issues of 'post-human security' discussed in subsection 3.3.

Besides cyber-attacks causing physical violence directly or indirectly, public actors tend to cyber-securitize when core issues of their national security, such as crucial confidential information, are at stake, which certainly may be interpreted as causing irreversible damage. The securitizing actions that have occurred in this context are 'exceptional measures' but not in the sense of involving the use of force. There is a significant degree of uncertainty regarding the adequate kind of response to cyber intelligence operations in this sense. On the one hand, NATO's Tallinn Manual states that "a state may not intervene, including by cyber means, in the internal or external affairs of another state" [53] and NATO does not exclude the possibility to respond kinetically to cyber-attacks [52]; on the other hand, international law tends to be rather permissive in regard to intelligence operations [73].

Attacked states are, therefore, aiming for targeted responses below the level of international conflict. For instance, the attack on the German Bundestag in 2015 brought about EU sanctions against the individuals and bodies involved within the EU's framework for a joint diplomatic response to malicious cyber activities [74]. The EU's so-called 'Cyber Diplomacy Toolbox' avoids national attribution [75]. This can be understood as an attempt to combine securitization with de-escalation. Another example of a response to intelligence operations is the US ban on Huawei, as recently renewed by Biden [76]. This also takes the conflict to the economic sphere without staging it as an international conflict and likewise combines de-escalation with securitization. Such approaches are certainly advisable from the perspective of my own JST framework, which is focused on subsidiarity.

More problematic are threats that surpass the traditional scope of cyber intelligence operations, particularly such threats that can be regarded as illegal interference with a state's self-determination, which can be understood as involving irreversible damage and can hardly be separated from a political understanding of international conflict. Self-determination is usually understood as regarding the constitution, i.e. the founding of a state. However, as Ohlin argues, this emphasis stems from an outdated state-centred understanding that ignores the continuous constitutive role of deliberative processes in society [77].

Cases such as the Russian meddling with the US general election, the Brexit referendum in 2016, and the French election campaign in 2017 have demonstrated that, due to their legitimization through deliberative discourses, democracies are particularly vulnerable to such interference as it undermines their own legitimacy. As a reaction to these threats, securitization occurred in the form of severe restrictions on the freedom

of speech. For instance, France passed the 'Loi Avia' (2020) against hate speech and 'LoiNo. 2018–1202' against disinformation [78]. Because it touched upon the core of French Republicanism, the law against disinformation was subjected to heated debate, and ultimately its constitutionality was mainly affirmed because it only applies during the three months prior to the elections. In turn, the law against hate speech was so clearly at odds with the liberal paradigm of free speech that it was partly revoked as unconstitutional.

Germany's Network Enforcement Act (NetzDG) from 2017, which served as an inspiration for the French legislation, regulates disinformation and hate speech [78]. In the research literature, the law is regularly criticized for incentivizing the over-blocking of content that is not clearly illegal [79]. Against the intentions of its creators, the law has inspired authoritarian and semi-authoritarian regimes, including Singapore, Russia, the Philippines, and Venezuela, which explicitly mentioned the German NetzDG as model legislation [80]. Interestingly, particularly in the case of the network enforcement act, nation states are not directly involved in securitization in a sense of limiting fundamental rights, but this task is outsourced to private actors that have purely economic motivations to engage in over-deletion. The Germany Director at Human Rights Watch argued that the law "turns private companies into overzealous censors to avoid steep fines" [81]. Another drastic example of the securitization of the digital public sphere is the unprecedented prohibition of the spread of Russian state-sponsored media in the EU in 2022 [82].

Using the framework of JST, the legitimacy of this kind of securitization of the digital public sphere seems highly questionable. Since democratic deliberation is to be 'protected' by these measures, it appears paradoxical that this is done by limiting the legal precondition of public deliberation, which is freedom of speech. Furthermore, securitization of the digital public sphere might work as a short-term remedy against cyber interference, but it is hardly adequate to avert this threat in a sustainable way. Floyd argues that desecuritization (which will be addressed in the next section) is an integral part of JST; according to her, securitization must have the aim of being reversed by desecuritization at one point, and it needs to include measures to aim for sustainable stability and "avoid renewed (...) securitization" [43].

In contrast, although these regulations restricting free speech online were implemented as a reaction to Russia's interference in 2016 and 2017 and the spread of war propaganda in 2022, there are no discussions regarding the reversal of these securitization processes. This is hardly surprising because this kind of securitization of the digital public sphere is evidently not sustainable, since it produces a version of the 'societal security dilemma' [4]: traditionally, this dilemma describes how, for instance in ethnic conflicts, states tend to strengthen their 'own' identity and, paradoxically, precisely this securitization move weakens these states' capacity to integrate minorities that do not identify with the main identity of these states. Therefore, securitization in this sense has a destabilizing effect.

Comparably, informational interference does not create new threats out of the blue, but it rather exploits already existing social fault lines and conflicts. Take for example the US's racial fault lines that were exploited

by Russia's support of White Supremacists and Black Lives Matter activists alike [83, 84]. Framing this kind of cyber interference as an international security issue shifts the focus away from resolving the domestic social tensions that create vulnerabilities to cyber interference in the first place. Therefore, whilst it satisfies the emotional need to identify a 'foreign' enemy, blaming social problems on 'Russian interference' can be expected to have a destabilizing effect in the long run.

Last but not least, securitizing the digital public sphere in this manner could have unintended ripple effects, since the formulation of prohibitions that cannot really be enforced undermines the credibility of the state (a realistic argument already used by Spinoza to make a point for freedom of speech [85]). For example, the ban on state-sponsored Russian media can be easily circumnavigated using VPNs. Furthermore, hate speech and disinformation will simply migrate to platforms such as Telegram that do not collaborate with the European authorities [86].

From a perspective focused on sustainable and subsidiary securitization, it would be more effective to stick to the EU's 'whole-of-society' approach and create sustainable resilience by promoting digital "media literacy as a key civic virtue" [64] Furthermore, the class divide contributes to the unquestioned acceptance of the claims of fake news [87]. Insofar as hostile cyber interference is exploiting social and other domestic fault lines, it would be more sustainable to tackle these vulnerabilities than to re-frame them as a result of foreign interference.

4.3. Private Companies as Norm Entrepreneurs and Securitizing Actors

Regarding private companies, it is debatable to what extent they can legitimately become securitizing actors. Evidently, private actors cannot engage in trade-offs between fundamental rights. When it comes to 'extraordinary measures' undertaken in securitization, all that private actors can do is allocate extraordinary resources to an issue or accept more friction in their services and processes due to security measures. This also means that companies will usually tend to engage less in cybersecurity if there are no direct financial concerns related to this (e.g. by protecting a company's trade secrets or customer data) or if they are not required by states to engage in cybersecurity or nudged by consumer demand to create products that offer a high degree of security.

Larger companies (such as Microsoft) are an exception to this rule: they seek cooperation with legislators and act as 'norm entrepreneurs' in matters related to cybersecurity, i.e., they engage in promoting societal, legal, and political norms regarding cybersecurity [88, 89]. Also, though the platforms involved in enforcing the securitization of the digital public sphere are not necessarily intending this, it serves to enhance their power, particularly regarding the over-deletion of content [90]. This points to an aspect of just war theory that Floyd failed to address in her JST. Classical just war theory argues that no financial or political gains should be connected with just war [91]. Likewise, no financial or political gains should be connected with just securitization. This restriction, which is important

to prevent just wars from being abused as justification for conquest or the acquisition of booty, is particularly important in relation to the involvement of private actors in securitization.

Companies large enough to afford compliance with complex cybersecurity regulations have a significant interest in promoting complex legislation since it can be a tool to push smaller players out of the market. To give some examples that do not directly relate to cybersecurity: in the field of data protection, the GDPR was hurting small and medium-sized enterprises (SMEs) more than bigger companies [92]. As a result of the learning derived from this asymmetrical effect, the new Digital Markets Act and the Digital Services Act (which likewise do not directly relate to cybersecurity) are exclusively regulating the activities of 'gatekeepers' and 'very large online platforms '(VLOPs) [93, 94]. Similar asymmetrical effects can be expected in regard to extended cybersecurity regulations if these do not exclusively target the bigger players. Therefore, the respective regulations should consider company size, and focus on subsidiarity in the sense that regulations should not overburden SMEs, thereby engendering market concentration. The EU's NIS Directive explicitly excludes small and micro enterprises for this reason [68].

Furthermore, the securitization of the digital public sphere by the restriction of freedom of speech gives large digital platforms a problematic degree of political power – even more problematic than the power that usually comes along with securitization, since private companies are not subject to the same legitimatory requirements as states [62]. Lehdonvirta argues that large digital platforms have transformed into 'Cloud Empires' because they guarantee and enforce social order and security on their virtual premises [95]. In summary, in contrast to the involvement of civil society stakeholders discussed in section 3, particularly the inclusion of Big Tech in the hyper-securitization of the digital public sphere is highly problematic since it can be associated with financial profits and gains in political power.

4.4. In favour of Self-Defence in Cyberspace

The issue of private self-defence in cyberspace is discussed in the literature [25, 62, 96]. Although this issue is situated below the threshold of physical violence, it is certainly a good example of securitization processes according to JST since it involves 'exceptional measures': the exemption from prosecution for acts that would normally be subject to this if they did not serve to avert a threat. In physical environments, a proportionate degree of physical resistance is justified if it serves to avert even non-violent acts of wrongdoing, for example, to deter threats to property, and if the act of resistance in question constitutes an adequate means of averting that threat, most notably in a pre-emptive sense. This is expressed in the stand-your-ground laws that are particularly far-reaching in the US [96].

These laws have been widely debated because they are connected with lethal violence and should be discussed critically [97]. Nevertheless, they express a fundamental principle of liberal legal and political philosophy: that individuals have the right to actively defend themselves if no other

remedy is available, following Grotius's notion of 'private just war', which I applied to justify self-defence in cyberspace in another article [25]. Most notably, laws regarding self-defence are rather focused on the aversion of the threat than on proportionality in the sense of retributive justice. Even in heavily gun-controlled Germany, it is possible to resort to physical defence against crimes that do not directly affect the body, for instance, theft [98].

In 2017 and 2019, the bipartisan 'Active Cyber Defence Certainty Act' proposal, known as the 'hack back' bill, was discussed in the US Congress [99]. It would have applied the principle of self-defence to cyberspace [100] by allowing private companies to engage in active self-defence against attackers. Interestingly, the bill applied the same distinction between irreversible and reversible damage on which Floyd bases the whole idea that securitization is permissible in the case of an existential threat. Directly complementary to Floyd's legitimization of 'exceptional measures' in cases in which threats cause irreversible damage, the 'hack back bill' restricts private legitimate self-defence to such cases in which this active self-defence "does not result in the destruction of data or result in an impairment of the essential operating functionality of the attacker's computer system" [99].

Of course, the so-called 'hack-back bill' has been widely criticized for enabling vigilantism [101]. However, it is not altogether apparent why the general right to self-defence should be virtually non-existent in cyberspace. Following the analogy to self-defence in offline environments, which can be aggressive as long as it clearly serves to avert a threat or constitutes a swift reaction to an attack, it is not even clear why these measures should be restricted to "defensive measures", as argued by Pattison [62]. It seems to be sufficient that they are occurring swiftly and are effective in averting or preventing a threat.

Whilst it may seem paradoxical, a greater degree of autonomous securitization on the subsidiary level of private actors might contribute to de-escalation and desecuritization, since it allows for a certain degree of cyber-securitization to occur on the level of immediate stakeholders without escalating to the national level. In this regard, securitization executed in a largely autonomous sense by private actors would be directly opposed to the highly problematic form of hyper-securitization involving an entanglement between Big Tech and governments as discussed earlier. Furthermore, particularly regarding cybercrime with an international dimension, for instance, cross-border economic espionage, keeping the whole conflict below the international level could contribute to de-escalation inasmuch as this subsidiary strategy has, at least, the potential to keep international actors out of relatively petty conflicts and to guarantee a certain degree of deterrence at the same time.

5. What is Just Termination regarding Cybersecurity?

5.1. Cybersecurity beyond the flawed distinction between 'Normal' and 'Exceptional'

Just termination is certainly one of the most interesting features of Floyd's JST. Just termination refers to desecuritization following the Copenhagen School approach, i.e., the reversal of securitization, "the shifting of issues out of emergency mode and into the normal bargaining processes of the political sphere" [7]. Most importantly, she sticks to her overall restrictive trajectory by emphasizing that whilst she denies the obligation to securitize (see section 3), she underlines that there is an obligation to de-securitize:

> Just desecuritization is about what desecuritizing actors are required to do, not about what such actors are permitted to [43].

Whereas Floyd is aware of the problems with the Copenhagen School's definition in this regard, her rationale remains that the desecuritized situation represents the 'normal' situation, in which civil liberties and fundamental rights are in full effect, whereas the securitized situation represents the situation in which these rights and liberties are partly suspended through 'exceptional measures'.

From an intercultural perspective, the Copenhagen School's definition is highly flawed, since it implicitly assumes that well-functioning liberal societies with a history of colonialist exploitation (in the style of Denmark in the 1990s) represent the 'normal' state of affairs (which they never did globally) [102]. Furthermore, this terminological-conceptual setup produces two significant problems. First, it does not apply to authoritarian regimes such as China, where the 'normal' situation is heavily securitized; second, it fails to consider the obvious fact that civil liberties and human rights do not represent a 'normal' state of affairs but are the product of the state's monopoly on violence, which could be understood as a form of continuous, low-intensity, day-to-day securitization (as argued in section 3).

Following the Copenhagen School's understanding of securitization, based on the dichotomy between the 'normal' and the 'exceptional', Floyd writes that "desecuritization of just securitization must occur when the initial and related new objective existential threats have been neutralized" [43]. This makes a good deal of sense, despite the problematic assumptions behind the dichotomy between securitization and desecuritization. Particularly since they involve trade-offs between fundamental rights, securitization measures need to be proportionate and then revoked once their necessity becomes less evident.

This understanding of securitization as an exceptional form of disruption also applies, in certain aspects, to cybersecurity. For instance, security concerns have been cited to legitimize exceptions to the WTO's free trade regime regarding the banning of Chinese 5G suppliers such as Huawei [103, 104]. Since these measures are exceptions, their justification based on national security concerns suggests that they should be reversed once the Chinese government ceases to constitute a threat. Generally speaking, desecuritization is crucial regarding digital technologies because they rely on a baseline situation characterized by the free flow of information and services.

However, precisely for this reason, the dichotomy between securitization and desecuritization is problematic regarding cybersecurity. As discussed in sections 2 and 3, cybersecurity is mostly connected to

everyday technological routines, rather than to 'exceptional measures'. Alongside this, due to the iterative nature of technological procedures, cybersecurity is not enacted momentarily and then reversed but is usually thought of as constituting a lasting feature that is inherently positive, as long as it guarantees the protection of human rights, such as the right to privacy. Rather than constituting an exceptional opposite to the free flow of information and services, cybersecurity must be combinable with openness, ideally with the greatest degree of openness. Cybersecurity, in the ideal typical sense (this means there are many exceptions to this), requires permanent securitization under the condition of permanent desecuritization.

5.2. Subsidiarity as structural Desecuritization

Precisely because of these non-dichotomous aspects discussed in the previous subsection, it is important to concentrate on desecuritization in relation to cybersecurity also. If there is no distinction between securitization and desecuritization, then anything goes. As already mentioned in section 4, the permanent character of cyber-securitization creates particular incentives for abuse. It can create lasting structures and a steady stream of revenue, which constitutes a strong incentive to engage in threat inflation and hyper-securitization within the "cyber-industrial complex" [11] or the "military-digital complex" [12]. The continuous securitization of the digital public sphere discussed in section 4 might be an example of the problematic and ambivalent outcome of this kind of securitization. In some cases, such as regarding the enticement to violence or slander, it may be understood as a protection of human rights; in other cases, that is when over-deletion occurs, it simply constitutes a restriction of the freedom of speech by which the state and corporations mutually enhance their powers and limit civil liberties.

Whilst acknowledging the importance of the temporal aspect of securitization and de-securitization, inasmuch as it provides a criterion to judge *the aims* of securitization and the degree to which it provides sustainable stability, a simple binary between the 'normal' de-securitized situation and the 'exceptional' securitized situation does not do justice to the complexity of the cybersecurity landscape, which involves a great diversity of actors, temporalities, and trajectories and must consider securitization and desecuritization in the same instance.

As discussed in sections 3 and 4, this complex situation can be, at least partially, resolved by a version of JST based on the principle of subsidiarity, which does not generally favour de-securitization but is more inclined to pursuing a form of structural de-securitization: whilst acknowledging the necessarily lasting character of low-intensity securitization on an everyday level, particularly if it guarantees human rights, such an approach avoids hyper-securitization by relating securitization to the stakeholders most directly affected and granting them far-reaching possibilities to participate actively and autonomously in securitization.

The risk of hyper-securitization may also exist in the private sector but due to the high costs related to security, besides Big Tech's 'Cloud Empires', market mechanisms generally counteract hyper-securitization.

Usually, private companies offer products with different levels of security, tailored to the need of individual customers and legal requirements. Turning securitization into a conscious consumer choice in this way is perhaps the best and most realistic road toward de-securitization since it moves securitization from the domain of the exceptional, instinctive, collective, and emotional to the domain of mundane and rational choices by mature individuals and their everyday security practices [24].

In contrast to this, as argued earlier, hyper-securitization primarily occurs within the framework of an "invisible handshake" [66] involving governments and Big Tech, which, in turn, should be observed critically, particularly regarding the acquisition of economic and political power through securitization.

6. Conclusion

In this contribution, I discussed the application of Floyd's JST to cybersecurity. As a reaction to the incompatibilities of JST and cybersecurity, I developed JST further to be more compatible with this specific socio-technical environment, which is characterized by the great importance of the private sector and the civilian nature of the digital public sphere. In general, I have strengthened human rights and the idea of subsidiarity, according to which executive measures should ideally be enacted by the lowest organizational level. The germ of my arguments can be found in Floyd's original JST, which restricts the legitimacy of securitization to such cases in which securitization can be reasonably expected to be successful in averting threats to the satisfaction of existential human needs.

Two of my adjustments to Floyd's original JST (fig. 1) are particularly crucial in relation to cybersecurity: first, nation states' role in cybersecurity cannot be adequately understood by assuming that they are only *permitted* to securitize, as this constitutes the focus of Floyd's original JST. Rather, since states are required to guarantee the human right to privacy, they do have *legal and moral obligations* to guarantee cybersecurity on a day-to-day basis. Regarding everyday security practices, opposing the Copenhagen School's questionable construction of a dichotomy between the 'exceptional' and the 'normal' situation, it is rather the maintenance of the 'normal' situation under the rule of law that requires quotidian low-intensity securitization involving the state's monopoly on violence. In this context, violence is rather to be understood as an implicit threat behind the state's regulatory function, than as something that is acted out in 'exceptional measures'. Indirectly, nation states' obligation to guarantee human rights in cyberspace also shapes the obligations of the private sector regarding cybersecurity. Moreover, the dichotomy between securitization and desecuritization is hardly applicable to cybersecurity since, in this context, securitization should occur under desecuritized conditions, guaranteeing an uninterrupted but secure global flow of information and services.

This non-dichotomous relationship between securitization and desecuritization is tackled by the principle of subsidiarity, which can be understood as a structural form of desecuritization. This principle does

justice to the central role of private actors regarding cybersecurity, which are often more competent in the identification of threats and the construction of adequate securitization and defence mechanisms than states are. The subsidiarity principle, as the central aspect of my account of JST developed in this paper, has a permissive and a restrictive aspect. On the one hand, it leads to the demand that stakeholders directly affected by a threat should have the opportunity to participate in decision-making processes regarding cyber-securitization; furthermore, these stakeholders should also be given the legal means to engage in active self-defence, as this is common in offline environments with stand-your-ground laws and similar rules in many liberal jurisdictions.

On the other hand, this focus on subsidiarity directly tackles the issue of threat inflation and hyper-securitization: hyper-securitization causes a 'cybersecurity dilemma' that poses a threat to international security if states choose to deter by threatening to defend forward and respond with kinetic means to cyber-attacks, or if they attempt to strengthen national security by submitting citizens to surveillance. Moreover, hyper-securitization produces a variant of the 'societal security dilemma' by shifting the focus away from sustainably resolving domestic social conflicts and attributing them to 'foreign' interference instead. As a general rule, such hyper-securitization in cyberspace is not likely to be caused by private actors alone but by public-private partnerships in the framework of the "invisible handshake" [66], including the "cyber-industrial complex" [11] or the "military-digital complex" [12].

This paper opens up a broad horizon for further research. Floyd's focus on 'human needs' could be discussed in more detail regarding its ethical foundations, particularly with respect to the upsides and downsides of effects-based, i.e. utilitarian, approaches and intention-based, i.e. deontological or rights-based, approaches. My discussion suggests that Floyd mixes these approaches in a problematic manner. This inserts elements of retributive justice into her JST that appear incompatible with her understanding of securitization as involving 'exceptional measures'. The whole issue of retributive justice regarding cybersecurity should be investigated further, including but not restricted to the difficult problem of how to relate judicial and moral categories focused on physical violence to cyberspace at all.

My approach to strengthening the liberal principle of subsidiarity in JST is certainly not without alternatives. Floyd's monograph on JST [43] and particularly her more open discussion of possible approaches to JST as "a meta-theoretical framework" [22] [20] will certainly produce many opportunities to address the details of possible applications of JST to cybersecurity. I believe that my focus on the principle of subsidiarity has resolved the basic problems of this application reasonably well, which consists of the incompatibility of day-to-day technological routines and the drastic securitization discourses of states as discussed in section 2. Nevertheless, I am fully aware of the fact that my endorsement of private active self-defence in cyberspace may appear highly problematic. But it seems the burden of proof falls on the side of those who argue that the right to self-defence, which is fundamental to liberal societies, should not apply in cyberspace, since this would constitute a form of 'cyberspace

exceptionalism' [105]. Similar to Floyd's assessment of her own JST, my application of JST to cyberspace represents just "one possible variant of such a theory" [22].

Funding

Johannes Thumfart received funding from Gerda Henkel Stiftung's special programme Security Society and the State and the European Union Horizon 2020 research programme under MSCA COFUND grant agreement 101034352 with co-funding from the VUB-Industrial Research Fund.

Acknowledgements

I would like to thank the two reviewers for their thoughtful comments that greatly improved the paper.

References

[1] J. P. Barlow. (2016, Jan. 20). *A declaration of the independence of cyberspace, Electronic Frontier Foundation*. [Online]. Available: https://www.eff.org/de/cyberspace-independence. [Accessed: July 1, 2021].

[2] R. Barbrook, A. Cameron, "The Californian ideology," *Science as Culture*, vol. 6, no. 1, pp. 44–72, 1996, doi: 10.1080/09505439609526455.

[3] M. Dunn Cavelty, "Cybersecurity between hypersecuritization and technological routine," in *Routledge handbook of international cybersecurity*, E. Tikk, M. Kerttunen, Eds. New York: Routledge, Taylor & Francis Group, 2020, pp. 11–21.

[4] J. Burton, C. Lain, "Desecuritising cybersecurity: towards a societal approach," *Journal of Cyber Policy*, vol. 5, no. 3, pp. 449–470, 2020, doi: 10.1080/23738871.2020.1856903.

[5] M. Lacy, D. Prince, "Securitization and the global politics of cybersecurity," *Global Discourse,* vol. 8, no. 1, pp. 100–115, 2018, doi: 10.1080/23269995.2017.1415082.

[6] L. Hansen, H. Nissenbaum, "Digital disaster, cyber security, and the Copenhagen school," *International Studies Quarterly*, vol. 53, no. 4, pp. 1155–1175, 2009.

[7] B. Buzan, O. Wæver, J. de Wilde, *Security: A new framework for analysis. Boulder*, Colo: Lynne Rienner Pub, 1998.

[8] J. L. Austin, *How to do things with words.* Oxford: Clarendon Press, 1962.

[9] C. Kinnvall, J. Mitzen, "Anxiety, fear, and ontological security in world politics: Thinking with and beyond Giddens," *International Theory*, vol. 12, no. 2, pp. 240–256, 2020, doi: 10.1017/S175297192000010X.

[10] R. McDermott, "Some emotional considerations in cyber conflict," *Journal of Cyber Policy*, vol. 4, no. 3, pp. 309–325, 2019, doi: 10.1080/23738871.2019.1701692.

[11] J. Brito, T. Watkins. (2012, Apr. 10). *Loving the cyber bomb? The dangers of threat inflation in cybersecurity policy, Mercatus Center*. [Online]. Available: https://www.mercatus.org/publications/technology-and-innovation/loving-cyber-bomb-dangers-threat-inflation-cybersecurity. [Accessed: July 2, 2021].

[12] R. W. McChesney, *Digital disconnect: How capitalism is turning the Internet against democracy*. New York: The New Press, 2013.

[13] V. Bernal, "The cultural construction of cybersecurity: Digital threats and dangerous rhetoric," *Anthropological Quarterly*, vol. 94, no. 4, pp. 611–638, 2021, doi: 10.1353/anq.2021.0037.

[14] M. C. Libicki, "Is there a cybersecurity dilemma?," *The Cyber Defense Review*, vol. 1, no. 1, pp. 129–140, 2016.

[15] M. Dunn Cavelty, "Breaking the cyber-security dilemma: Aligning security needs and removing vulnerabilities," *Science and Engineering Ethics*, vol. 20, no. 3, pp. 701–715, 2014, doi: 10.1007/s11948-014-9551-y.

[16] B. C. Taylor, "Defending the state from digital deceit: The reflexive securitization of deepfake," *Critical Studies in Media Communication*, vol. 38, no. 1, pp. 1–17, 2021, doi: 10.1080/15295036.2020.1833058.

[17] N. Kshetri, *The quest to cyber superiority: Cybersecurity regulations, frameworks, and strategies of major economies.* New York: Springer, 2016.

[18] P. Roe, *Ethnic violence and the societal security dilemma*. London, New York: Routledge, 2005.

[19] R. Floyd, *The morality of security: A theory of just securitization*. New York: Cambridge University Press, 2019.

[20] R. Floyd, "The promise of theories of just securitization," in *Ethical security studies: A new research agenda*, J. Nyman, A. Burke, Eds. Abingdon, Oxon, New York: Routledge, 2016, pp. 75–88.

[21] R. Floyd, "Can securitization theory be used in normative analysis? Towards a just securitization theory," *Security Dialogue*, vol. 42, no. 4–5, pp. 427–439, 2011, doi: 10.1177/0967010611418712.

[22] R. Floyd, "States, last resort, and the obligation to securitize," *Polity*, vol. 51, no. 2, pp. 378–394, 2019, doi: 10.1086/701886.

[23] J. Thumfart, "The norm development of digital sovereignty between China, Russia, the EU and the US: From the late 1990s to the Covid-crisis 2020/21 as catalytic event," in *Enforcing rights in a changing world*, D. Hallinan, R. Leenes, P. de Hert, Eds. London: Hart Publishing, 2021, pp. 1–44.

[24] M. Dunn Cavelty, F. J. Egloff, "Hyper-securitization, everyday security practice and technification: Cyber-security logics in Switzerland," *Swiss Political Science Review*, vol. 27, no. 1, pp. 139–149, 2021, doi: 10.1111/spsr.12433.

[25] J. Thumfart, "Public and private just wars: Distributed cyber deterrence based on Vitoria and Grotius," *Internet Policy Review*, 2020, doi: 10.14763/2020.3.1500.

[26] G. Dimari, N. Papadakis, "The securitization of the Covid-19 pandemic in Greece: A just or unjust securitization?," *Quality & Quantity*, 2022, doi: 10.1007/s11135-022-01341-9.

[27] A. C. Dwyer, C. Stevens, L. P. Muller, M. D. Cavelty, L. Coles-Kemp, P. Thornton, "What can a critical cybersecurity do?," *International Political Sociology*, vol. 16, no. 3, p. olac013, 2022, doi: 10.1093/ips/olac013.

[28] T. Maurer, *Cyber mercenaries: The state, hackers, and power. Cambridge*, New York, Port Melbourne, New Delhi, Singapore: Cambridge University Press, 2018.

[29] E. Lilli, "Redefining deterrence in cyberspace: Private sector contribution to national strategies of cyber deterrence," *Contemporary Security Policy*, vol. 42, no. 2, pp. 163–188, 2021, doi: 10.1080/13523260.2021.1882812.

[30] S. Haataja, *Cyber attacks and international law on the use of force: The turn to information ethics*. Abingdon, Oxon, New York: Routledge, 2019.

[31] C. J. Finlay, "Just war, cyber war, and the concept of violence," *Philosophy & Technology*, vol. 31, no. 3, pp. 357–377, 2018, doi: 10.1007/s13347-017-0299-6.

[32] F. J. Egloff, J. Shires, "Offensive cyber capabilities and state violence: Three logics of integration," *Journal of Global Security Studies*, vol. 7, no. 1, p. ogab028, 2021, doi: 10.1093/jogss/ogab028.

[33] C. Schmitt, Political theology: Four chapters on the concept of sovereignty. Chicago: University of Chicago Press, 2005.

[34] J. Habermas, *The structural transformation of the public sphere: an inquiry into a category of bourgeois society*. Cambridge: MIT press, 1992.

[35] C. Schmitt, "The concept of the political," in *The concept of the political*, G. Schwab, Ed. Chicago: University of Chicago Press, 2007, pp. 19–79.

[36] J. P. Burgess, N. Mouhleb. (2007). *A presentation of the state of societal security in Norway, PRIO International Peace Research Institute*, Oslo. [Online]. Available: https://www.prio.org/publications/7197. [Accessed: July 2, 2021].

[37] S. Elbe, "Should HIV/AIDS be securitized? The ethical dilemmas of linking HIV/AIDS and security," *International Studies Quarterly*, vol. 50, no. 1, pp. 119–144, 2006, doi: 10.1111/j.1468-2478.2006.00395.x.

[38] N. J. Jackson, "International organizations, security dichotomies and the trafficking of persons and narcotics," in "Post-soviet central Asia: A critique of the securitization framework," *Security Dialogue*, vol. 37, no. 3, pp. 299–317, 2006.

[39] F. Robinson, "Feminist care ethics and everyday insecurities," in *Ethical security studies: A new research agenda*, J. Nyman, A. Burke, Eds. Abingdon, Oxon, New York: Routledge, 2016, pp. 116–130.

[40] M. Foucault, *The archaeology of knowledge*. New York: Vintage Books, 2010.

[41] L. Tien, "Publishing software as a speech act," *Berkeley Technology Law Journal*, vol. 15, no. 2, pp. 629–712, 2000.

[42] A. R. Galloway, *Protocol: How control exists after decentralization. Cambridge*, Mass: MIT Press, 2004.

[43] R. Floyd, *The morality of security: A theory of just securitization*. New York: Cambridge University Press, 2019.

[44] S. Ghafur, S. Kristensen, K. Honeyford, G. Martin, A. Darzi, P. Aylin, "A retrospective impact analysis of the WannaCry cyberattack on the NHS," *npj Digital Medicine*, vol. 2, no. 1, p. 98, 2019, doi: 10.1038/s41746-019-0161-6.

[45] J. Waldron, "The rule of international law," *Harvard Journal of Law and Public Policy*, vol. 30, no. 1, pp. 15–30, 2006.

[46] R. J. Deibert, "Toward a human-centric approach to cybersecurity," *Ethics & International Affairs*, vol. 32, no. 4, pp. 411–424, 2018, doi: 10.1017/S0892679418000618.

[47] N. Möllers, "Making digital territory: Cybersecurity, techno-nationalism, and the moral boundaries of the state," *Science, Technology, & Human Values*, vol. 46, no. 1, pp. 112–138, 2021, doi: 10.1177/0162243920904436.

[48] M. Mueller, *Will the Internet fragment? Sovereignty, globalization and cyberspace*. Cambridge, United Kingdom, Malden: Polity Press, 2017.

[49] A. Bradford. (2020). *The Brussels effect: How the European Union rules the world, Oxford University Press*. [Online]. Available: https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190088583.001.0001/oso-9780190088583. [Accessed: Feb. 13, 2022].

[50] M. S. Erie, T. Streinz. (2021). "The Beijing effect: China's digital silk road as transnational data governance," *New York University Journal of International Law and Politics*, vol. 54, no. 1. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3810256. [Accessed: Feb. 13, 2022].

[51] A. Kokas, *Trafficking data: How China is winning the battle for digital sovereignty*. New York: Oxford University Press, 2022, doi: 10.1093/oso/9780197620502.001.0001.

[52] M. Prucková, *Cyber attacks and Article 5 – a note on a blurry but consistent position of NATO, NATO Cooperative Cyber Defence Centre of Excellence*. [Online]. Available: https://ccdcoe.org/library/publications/cyber-attacks-and-article-5-a-note-on-a-blurry-but-consistent-position-of-nato/. [Accessed: Nov. 2, 2022].

[53] M. N. Schmitt, *NATO Cooperative Cyber Defence Centre of Excellence, Eds., Tallinn manual 2.0 on the international law applicable to cyber operations, 2nd ed. Cambridge*, United Kingdom, New York, USA: Cambridge University Press, 2017.

[54] F. J. Egloff, "Contested public attributions of cyber incidents and the role of academia," *Contemporary Security Policy*, vol. 41, no. 1, pp. 55–81, 2020, doi: 10.1080/13523260.2019.1677324.

[55] W. Cong, J. Thumfart, "A Chinese precursor to the digital sovereignty debate: Digital anti-colonialism and authoritarianism from the Post–Cold War Era to the Tunis Agenda," *Global Studies Quarterly*, vol. 2, no. 4, 2022, doi: https://doi.org/10.1093/isagsq/ksac059.

[56] U. Nations. (1948, Dec. 10). *Universal Declaration of Human Rights.* [Online]. Available: https://www.un.org/en/about-us/universal-declaration-of-human-rights. [Accessed: Nov. 14, 2022].

[57] M. Michaelsen, J. Thumfart, "Drawing a line: Digital transnational repression against political exiles and host state sovereignty," *European Journal of International Security*, pp. 1–21, 2022, doi: 10.1017/eis.2022.27.

[58] A. Mitchell, "Posthuman security / ethics," in *Ethical security studies: A new research agenda*, J. Nyman, A. Burke, Eds. Abingdon, Oxon, New York: Routledge, 2016, pp. 60–72.

[59] A. Avery. (2020). *Cybersecurity Scenario Modeling: Imagining the Black Swans for Digital Infrastructures Risk Management.* [Online]. Available: https://aisel.aisnet.org/sais2020/5. [Accessed: Nov. 14, 2021].

[60] I.-C. Tsai, "Flash crash and policy uncertainty," *Journal of International Financial Markets, Institutions and Money*, vol. 57, pp. 248–260, 2018, doi: 10.1016/j.intfin.2018.09.002.

[61] K. Bannelier, T. Christakis. (2017). *Cyber-Attacks – prevention-reactions: The role of states and private actors, Les Cahiers de la Revue Défense Nationale, Paris*. [Online]. Available: https://ssrn.com/abstract=2941988. [Accessed: Nov. 14, 2021].

[62] J. Pattison, "From defence to offence: The ethics of private cybersecurity," *European Journal of International Security*, vol. 5, no. 2, pp. 233–254, 2020, doi: 10.1017/eis.2020.6.

[63] P. Beuth, J. Breithut. (2021, Sep. 12). *40 Jahre CCC: Chaos macht Politik, Der Spiegel*. [Online]. Available: https://www.spiegel.de/netzwelt/netzpolitik/40-jahre-ccc-chaos-macht-politik-a-655ecc5b-d135-4ae5-846e-535d340448c3. [Accessed: Aug. 30, 2022].

[64] M. Wigell, H. Mikkola, T. Juntunen. (2021). *Best practises in the whole of society approach in countering hybrid threats*. [Online]. Available: https://www.europarl.europa.eu/committees/de/best-practises-in-the-whole-of-society-a/product-details/20210531CAN61132. [Accessed: July 1, 2021].

[65] *Cybersecurity & Infrastructure Security Agency, Joint cyber defense collaborative*. [Online]. Available: https://www.cisa.gov/jcdc. [Accessed: Aug. 30, 2022].

[66] M. D. Birnhack, N. Elkin-Koren, "The invisible handshake: The reemergence of the state in the Digital Environment," *SSRN Electronic Journal*, 2003, doi: 10.2139/ssrn.381020.

[67] M. Michaelsen. (2020). *The digital transnational repression toolkit, and its silencing effects, Freedom House*. [Online]. Available: https://freedomhouse.org/report/special-report/2020/digital-transnational-repression-toolkit-and-its-silencing-effects. [Accessed: May 29, 2021].

[68] European Union. (2016). Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2016/1148/oj. [Accessed: Nov. 4, 2022].

[69] *The Clean Network. (2017–2021). United States Department of State*. [Online]. Available: https://2017-2021.state.gov/the-clean-network/. [Accessed: Sep. 2, 2022].

[70] G. Maihold. (2022). *A new geopolitics of supply chains: The rise of friend-shoring, Stiftung Wissenschaft und Politik*. [Online]. Available: https://www.swp-berlin.org/10.18449/2022C45/. [Accessed: Oct. 26, 2022].

[71] F. J. Egloff, J. Shires, "The better angels of our digital nature? Offensive cyber capabilities and state violence," *European Journal of International Security*, pp. 1–20, 2021, doi: 10.1017/eis.2021.20.

[72] *CBS News. (2008, May 8). Epilepsy site hacked with seizure images*. [Online]. Available:

https://www.cbsnews.com/news/epilepsy-site-hacked-with-seizure-images/. [Accessed: May 17, 2022].

[73] A. Deeks, "Confronting and adapting: intelligence agencies and international law," *Virginia Law Review*, vol. 102, no. 3, pp. 599–685, 2016.

[74] *European Council, Malicious cyber-attacks: EU sanctions two individuals and one body over 2015 Bundestag hack*. [Online]. Available: https://www.consilium.europa.eu/en/press/press-releases/2020/10/22/malicious-cyber-attacks-eu-sanctions-two-individuals-and-one-body-over-2015-bundestag-hack/. [Accessed: Nov. 6, 2022].

[75] F. Dumortier, V. Papakonstantinou, P. de Hert. (2020, Sep. 28). *EU sanctions against cyber-attacks and defense rights: Wanna Cry?, European Law Blog*. [Online]. Available: https://europeanlawblog.eu/2020/09/28/eu-sanctions-against-cyber-attacks-imposed-and-defense-rights-wanna-cry/. [Accessed: July 20, 2022].

[76] *BBC News. (2021, Nov. 12). US President Joe Biden tightens restrictions on Huawei and ZTE*. [Online]. Available: https://www.bbc.com/news/technology-59262329. [Accessed: Nov. 5, 2022].

[77] J. D. Ohlin. (2017). *Did Russian cyber interference in the 2016 election violate international law?, Texas Law Review, vol. 95, no. 7* [Online]. Available: https://texaslawreview.org/russian-cyber-interference-2016-election-violate-international-law/. [Accessed: June 30, 2022].

[78] D. Steiger, "Protecting democratic elections against online influencevia 'fake news' – and hate speech – the french Loi Avia and Loi No. 2018–1202, the GermanNetwork enforcement act and the EU's Digital Services actin light of the right to freedom of expression," in *Theory and practice of the European Convention on Human Rights*, S. Schiedermair, A. Schwarz, D. Steiger, Eds. Baden-Baden: Nomos, 2022, pp. 165–214.

[79] L. Marc, *Das NetzDG in der praktischen Anwendung: Eine Teilevaluation des Netzwerkdurchsetzungsgesetzes. Carl Grossmann*, 2021. doi: 10.24921/2021.94115953.

[80] J. Mchangama, J. Fiss. (2019). *The digital Berlin Wall: How Germany (accidentally) created a prototype for global online censorship, Justitia, Copenhagen.* [Online]. Available: https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2019/11/Analyse_The-Digital-Berlin-Wall-How-Germany-Accidentally-Created-a-Prototype-for-Global-Online-Censorship.pdf. [Accessed: Nov. 5, 2022].

[81] Human Rights Watch. (2018, Feb. 14). *Germany: Flawed social media law*. [Online]. Available: https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law. [Accessed: Sep. 8, 2022].

[82] B. Baade. (2022, Mar. 8). *The EU's 'Ban' of RT and Sputnik, Verfassungsblog*. [Online]. Available: https://verfassungsblog.de/the-eus-ban-of-rt-and-sputnik/. [Accessed: Apr. 6, 2022].

[83] S. Bradshaw, R. DiResta, C. Miller, "Playing both sides: Russian state-backed media coverage of the #blacklivesmatter movement," *The International Journal of Press/Politics*, 2022, doi: 10.1177/19401612221082052.

[84] T. Snyder, The road to unfreedom: Russia, Europe, America. New York: Tim Duggan Books, 2018.

[85] B. de Spinoza, Theological-political treatise. Cambridge, New York: Cambridge University Press, 2007.

[86] S. Krempl. (2022, Mar. 24). NetzDG-Streit mit Telegram: Deutsche Justiz wendet Zustellungstrick an, heise online. [Online]. Available: https://www.heise.de/news/NetzDG-Streit-mit-Telegram-Deutsche-Justiz-wendet-Zustellungstrick-an-6624629.html. [Accessed: Sep. 8, 2022].

[87] D. A. Scheufele, N. M. Krause, "Science audiences, misinformation, and fake news," Proceedings of the National Academy of Sciences, vol. 116, no. 16, pp. 7662–7669, 2019, doi: 10.1073/pnas.1805871115.

[88] L. M. Hurel, L. C. Lobato, "Unpacking cyber norms: private companies as norm entrepreneurs," Journal of Cyber Policy, vol. 3, no. 1, pp. 61–76, 2018, doi: 10.1080/23738871.2018.1467942.

[89] C. M. Glen, "Norm entrepreneurship in global cybersecurity," Politics & Policy, vol. 49, no. 5, pp. 1121–1145, 2021, doi: 10.1111/polp.12430.

[90] R. Caplan, "The Artisan and the Decision Factory: The organizational dynamics of private speech governance," in Digital technology and democratic theory, L. Bernholz, H. Landemore, R. Reich, Eds. Chicago: University of Chicago Press, 2020, pp. 167–190.

[91] J. Thumfart, "Francisco de Vitoria and the Nomos of the Code: The Digital Commons and Natural Law, digital communication as a human right, just cyber-warfare," in At the origins of modernity, vol. 10, J. M. Beneyto, J. Corti Varela, Eds. Cham: Springer International Publishing, 2017, pp. 197–217.

[92] D. Barnard-Wills, L. Cochrane, K. Matturi, F. Marchetti. (2019). Report on the SME experience of the GDPR, Trilateral Research, Budapest – Brussels – Waterford, STAR II Deliverable D2.2. [Online]. Available: https://www.trilateralresearch.com/wp-content/uploads/2020/01/STAR-II-D2.2-SMEs-experience-with-the-GDPR-v1.0-.pdf. [Accessed: Nov. 5, 2022].

[93] European Parliament. (2022, Mar. 24). Deal on digital markets act: Ensuring fair competition and more choice for users. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20220315IPR25504/deal-on-digital-markets-act-ensuring-fair-competition-and-more-choice-for-users. [Accessed: Apr. 14, 2022].

[94] European Parliament. (2020). Digital Services Act – questions and answers. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348. [Accessed: Apr. 14, 2022].

[95] V. Lehdonvirta, Cloud empires: How digital platforms are overtaking the state and how we can regain control. Cambridge, Massachusetts: The MIT Press, 2022.

[96] H. Gandhi, "Active cyber defense certainty: A digital self-defense in the modern age," Oklahoma City University Law Review, vol. 43, pp. 101–131, 2019.

[97] C. McClellan, E. Tekin, "Stand your ground laws, homicides, and injuries," Journal of Human Resources, vol. 52, no. 3, pp. 621–653, 2017, doi: 10.3368/jhr.52.3.0613-5723R2.

[98] J. Bülte, "Zur Verhältnismäßigkeit der Notwehr und Art. 103 Abs. 2 GG als Schranken-Schranke," Neue Kriminalpolitik, vol. 28, no. 2, pp. 172–192, 2016.

[99] T. Graves. (2019, June 28). Text – H.R.3270 – 116th Congress (2019–2020): Active Cyber Defense Certainty Act. [Online]. Available: https://www.congress.gov/bill/116th-congress/house-bill/3270/text. [Accessed: July 1, 2021].

[100] M. Noone. (2018, Feb. 2). Self-defense goes cyber: Congress considers a bill permitting victims of cyberattacks to 'hack back', University of Baltimore Law Review. [Online]. Available: https://ubaltlawreview.com/2018/02/02/self-defense-goes-cyber-congress-considers-a-bill-permitting-victims-of-cyberattacks-to-hack-back/. [Accessed: Sep. 4, 2022].

[101] M. Giles. (2019). Five reasons 'hacking back' is a recipe for cybersecurity chaos, MIT Technology Review. [Online]. Available: https://www.technologyreview.com/2019/06/21/134840/cybersecurity-hackers-hacking-back-us-congress/. [Accessed: Sep. 10, 2022]

[102] A. Howell, M. Richter-Montpetit, "Is securitization theory racist? Civilizationism, methodological whiteness, and antiblack thought in the Copenhagen school," *Security Dialogue*, vol. 51, no. 1, pp. 3–22, 2020, doi: 10.1177/0967010619862921.

[103] S.-y. Peng, "Cybersecurity threats and the WTO national security exceptions," *Journal of International Economic Law*, vol. 18, no. 2, pp. 449–478, 2015, doi: 10.1093/jiel/jgv025.

[104] S. Nebehay. (2020, June 11). *China hits back at U.S. telecom supply chain order at WTO, Reuters*. [Online]. Available: https://www.reuters.com/article/us-usa-trade-china-wto-idUSKBN23I32V. [Accessed: Nov. 2, 2022].

[105] J. Cohen. (2007). *Cyberspace as/and space, Georgetown Law Faculty Publications and Other Works*. [Online]. Available: https://scholarship.law.georgetown.edu/facpub/807. [Accessed: Nov. 2, 2022].

Towards an Efficient and Coherent Regulatory Framework on Cybersecurity in the EU: The Proposals for a NIS 2.0 Directive and a Cyber Resilience Act

Sandra Schmitz-Berndt | **Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg**, ORCID: 0000-0001-9443-9206

Mark D. Cole | **Faculty of Law, Economics and Finance, Department of Law, University of Luxembourg, Luxembourg**, ORCID: 0000-0003-4382-8791

Abstract

Cybersecurity regulation in the EU has long been implemented in a piece-meal fashion resulting in a fragmented regulatory landscape. Recent developments triggered the EU to review its approach which has not resulted in the envisaged high level of cyber resilience across the Union. The paper addresses the EU's limited mandate to regulate cybersecurity and outlines how the internal market rationale serves as a basis to harmonise cybersecurity legislation in the EU Member States. In that regard, the recent Proposal for a NIS 2.0 Directive (adopted by the European Parliament in November 2022) and the Proposal for a Cyber Resilience Act (published in September 2022) highlight how the EU seeks to align legislation and reduce complexity between different, often sectoral regulatory approaches to cybersecurity, while at the same time extending regulation in a view to achieve a high level of cybersecurity across the EU. As regards the latter, the paper also outlines how the Cyber Resilience Act will complement the NIS 2.0 Directive in order to close existing regulatory gaps.

Corresponding author:
Sandra Schmitz-Berndt, Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, 6, avenue de la Fonte, L-4363 Esch-sur-Alzette, Luxembourg; ORCID: 0000-0001-9443-9206; sandra.schmitz@uni.lu

Keywords
Cyber Resilience Act, cybersecurity, EU legislative framework, NIS 2.0 directive

Cite this article as: S. Schmitz-Berndt, M.D. Cole, "Towards an Efficient and Coherent Regulatory Framework on Cybersecurity in the EU: The Proposals for a NIS 2.0 Directive and a Cyber Resilience Act," ACIG, vol. 1, no. 1, pp. 148–168, 2022,
DOI: 10.5604/01.3001.0016.1323

1. Introduction

Cybersecurity threats concern every legal entity and every natural person in our society. With digital transformation and interconnectedness of society, network and information systems (NIS) have developed into an essential commodity in everyday life. The COVID-19 pandemic triggered a shift to remote working with a surge in connections from private to corporate systems and an unprecedented adoption of telecommuting and video conferencing. With telework becoming the norm in many sectors and industries, many corporate networks became more vulnerable to cyberattacks. At the same time, physical and digital infrastructures are increasingly interconnected and interdependent. Also, different services and sectors of our economies are interconnected and are growing more dependent on NIS than ever before. Apart from the economic aspect, the speedy digital transformation also means that our society is more interconnected.

The unprecedented digital dependencies that we see today mean that there is to an increased attack surface posing numerous challenges of managing cybersecurity [1]. The steady increase in the number of users and connections also creates new vulnerabilities [2]. New opportunities arise for cyber-dependent crime. Not surprisingly, within the last months, a notable increase in the number of cyberattacks on citizens, businesses and critical infrastructures has been reported [2] including for instance ransomware attacks on health services [3, 4], and on public administration [5]. In 2021, Germany faced a 360% increase of such ransomware attacks [4]. Cyberattacks also targeted a range of EU institutions, including the European Commission, the European Medicines Agency and the European Banking Authority [6]. Earlier large scale cyber espionage campaigns on agencies and ministries across the European states targeted for instance the Norwegian Parliament [7], the German Parliament and the federal government's internal communications network [8], and a French software form which supplies the French Ministry of Justice [9]. There is sufficient evidence that the number, magnitude, sophistication, frequency and impact of cybersecurity incidents are increasing, and that this presents a major threat to the functioning of network and information systems (NIS). A disruption in one state can have cascading effects with ramifications in numerous other states.

Furthermore, renewed geopolitical tension between the West, Russia and China, and ultimately Russia's war of aggression against Ukraine have proven that the resilience of EU critical infrastructures is at risk from both physical and cyber threats [10]. This has only recently

been highlighted by for instance the sabotage of the Nord Stream gas pipelines [11], the German rail network [12] and the cyberattack on the U.S. telecommunications company Viasat which was launched in parallel to the physical invasion of Ukraine and that affected customers across Europe [13].

Against that background, states and also the EU are becoming very active to strengthen the physical and cyber resilience with the latest effort being a Proposal for a Commission Recommendation to strengthen the resilience of critical infrastructures [14] in October 2022. Also, in June 2022, a political agreement [15] has been reached on the Proposal for Directive on the resilience of critical entities [16], which seeks to revise the current approach to critical infrastructure protection taken under the European Critical Infrastructures Directive[1]. Apart from legislative activities in the area of physical security, in the area of cybersecurity, the EU Commission [10] stresses the need for the application of an updated and comprehensive legal framework to be accelerated in order to strengthen cyber resilience, while at the same time striving to become a leader in cybersecurity [17]. As such, cybersecurity has been a top priority of the EU Commission since the first cybersecurity strategy in 2013 [18], which marked the formal establishment of 'cybersecurity' as a new policy area; followed by the Digital Single Market Strategy for Europe [19], where the digitalisation of the internal market is characterised by a high degree of trust, security, safety and choice for consumers. In 2022, significant steps have been taken to advance the regulation of cybersecurity: most importantly, following a political agreement [20] on the Proposal for a new NIS Directive (NIS 2.0 Proposal) [21] in May 2022[2], the European Parliament adopted a consolidated text of the Proposal [22] in a first reading on 10 November 2022. The NIS 2.0 Directive will replace the existing 2016 NIS Directive[3]. In the same parliamentary session, the European Parliament adopted a consolidated text for a Regulation on digital operational resilience for the financial sector (DORA Regulation) [23], which seeks to strengthen the IT security of financial entities. Cybersecurity is also subject to two proposals aiming to boost cybersecurity and information security in EU institutions, bodies, offices and agencies [24, 25][4] of March 2022, and the Proposal for a Cyber Resilience Act (CRA Proposal) [26] of September 2022.

This paper will first provide an introduction into the regulation of cybersecurity in the EU in general, in particular into the EU mandate to regulate cybersecurity (section 2), before it will address in detail how the NIS 2.0 Directive and the CRA seek to improve the overall cybersecurity across the EU (section 3.). The focus on the NIS 2.0 Directive and the CRA is owed to the fact that both instruments regulate cyber aspects of ICT horizontally instead of introducing different, sectoral regulatory approaches to cybersecurity: the NIS 2.0 Directive addresses specific services based on digital infrastructures, while the CRA addresses the underlying technology of digital products and ancillary services. Section 3 also addresses how the NIS 2.0 Directive and the CRA reflect a risk-based approach to technology regulation and how they complement each other.

1 — Council Directive 2008/114/EC of 08.12.2008 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection, OJ L 345, 23.12.2008, p. 75. The ECI Directive only applies to the energy and transport sectors.

2 — The NIS 2.0 Proposal mirrors the approach taken by the aforementioned Proposal for a CER Directive for the cyber dimension of the services covered; matters covered by the NIS 2.0 Directive will be excluded from the scope of the CER Directive.

3 — Directive (EU) 2016/1148 of the European Parliament and of the Council of 06.07.2016 concerning measures for a high common level of security of network and information systems across the Union, OJ L 194, 19.07.2016, p. 1.

4 — The Proposal for a Regulation laying down measures on a high level of cybersecurity at the institutions, bodies, offices and agencies of the Union will put in place a framework for governance, risk management and control in the cybersecurity field. The Regulation will also extend the mandate of CERT-EU. The Proposal for a Regulation on information security in the institutions, bodies, offices and agencies of the Union will create a minimum set of information security rules and standards for all EU institutions, bodies, offices and agencies to ensure an enhanced and consistent protection against evolving threats.

2. Cybersecurity Regulation in the EU in Generaln

2.1. Cybersecurity as a EU Policy Field

The EU's approach to cybersecurity policy is mainly addressed in the EU Cybersecurity Strategies. The first EU Cybersecurity Strategy [18] of February 2013 represented the EU's comprehensive vision on how to best prevent and respond to cyber disruptions and attacks while at the same time furthering European values of freedom and democracy and ensuring the digital economy can safely grow. The Strategy also provided – although only in a footnote – a definition of cybersecurity as cybersecurity commonly referring 'to the safeguards and actions that can be used to protect the cyber domain, both in the civilian and military fields, from those threats that are associated with or that may harm its interdependent networks and information infrastructure' [18]. Accordingly, the primary objectives of cybersecurity were identified as preserving 'the availability and integrity of the networks and infrastructure and the confidentiality of the information contained therein' [18][5].

As outlined in the introduction, with increased interconnection new challenges arose accompanied by growing concerns about the privacy and security of businesses and individuals in cyberspace. The WannaCry, Petya and NotPetya ransomware attacks in 2017 proved that cyberattacks are the new reality, and perfectly highlighted the cascading effects that may affect more entities than anticipated [27]. In response to the attack, the European Commission and the High Representative of the Union for Foreign Affairs and Security Policy published a new Cybersecurity Strategy [28] in 2017. The 2017 Cybersecurity Strategy highlighted the need for measures that would allow building greater EU resilience to cyberattacks, facilitating their detection, and strengthening international cooperation on cybersecurity. The two Cybersecurity Strategies resulted in legislation, namely the NIS Directive in 2016 (as a result of the 2013 Cybersecurity Strategy), which was the first EU-wide legislation on cybersecurity, and the Cybersecurity Act (CSA)[6] in 2019 (as a result of the 2017 Cybersecurity Strategy), which strengthens the role and mandate of the European Union Agency for Network and Information Security (ENISA) and introduces the legal basis to adopt an EU-wide cybersecurity certification scheme for ICT products. Several soft law instruments complemented these regulatory initiatives, for instance a Recommendation on the cybersecurity of 5G networks [29]. Also, strategic investments in digital capacity and infrastructure building took place. With reducing cybercrime also being a policy aim of the Cybersecurity Strategies, legislative and policy measures were taken forward in judicial and law enforcement matters, for instance the Directive on Attacks against Information Systems[7]. Cybersecurity has also been a driver for security and defence integration in the EU [30].

Building on the Commission Communication Shaping Europe's digital future [31] and the EU Security Union Strategy [32], the European Commission published a new Cybersecurity Strategy [33] in 2020 accompanied by a Proposal for a NIS 2.0 Directive [21] and a Proposal for Directive on the resilience of critical entities [16]. The 2020 Strategy pays regard to the speed of digital transformation in a complex threat

5 —— This definition deviated to some extent from a previous suggestion by ENISA, see on this [27

6 —— Regulation (EU) 2019/881 of the European Parliament and of the Council of 17.04.2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act), OJ L 151, 07.06.2019, p. 15.

7 ——Directive 2013/40/EU of the European Parliament and of the Council of 12.08.2013 on attacks against information systems and replacing Council Framework decision 2005/222/JHA, OJ L 218, 14.08.2013, p. 8.].

environment, which is compounded by geopolitical tensions over the global and open Internet and over control of technologies across the supply chain. The main objectives of the Strategy are (1) resilience, technological sovereignty and leadership, (2) building operational capacity to prevent, deter and respond, and (3) advancing a global and open cyberspace. The short interval between the 2017 and 2020 Strategies reflects the political acquis that there is an urgent need for action; as already addressed in the introduction, the speed of regulatory action is accelerating.

2.2. The EU's Limited Mandate to Regulate Cybersecurity

A fundamental principle of EU law is the principle of conferral under which the EU acts only within the limits of the competences conferred upon it by the Member States. In general, the EU can legislate in areas where it is more appropriate for the EU to act than for the Member States individually. The introduction of any regulatory measure at EU level requires a legal basis. For cybersecurity, the EU Treaties do not provide such a unifying legal basis. Moreover, if one considers cybersecurity as part of national security, Article 4(2) TFEU provides that national security remains the sole responsibility of each Member State. Cyber policy, especially in the context of the protection of critical infrastructures has a national security dimension [34]. However, the cybersecurity dimension goes beyond national security, cybersecurity also has cross-border effects. What is more, not all cybersecurity aspects fall outside the scope of EU law: there are policy domains which are affected by cyber threats and in which the Treaties do confer powers upon the EU.

The EU's regulatory approach towards internet and cyberspace has long been focusing on economic growth under the single market rationale. Under this rationale, the EU deploys its political and legal mandate to regulate the internal market to issue common policies and legislation on cybersecurity. The legal basis for this is Article 114 TFEU[8], which provides a very versatile legislative basis for the issuance of legislation that serves the aim of smoothening the functioning of the internal market. By establishing a link between cybersecurity and the smooth functioning of the internal market, the European Commission provided a justification for acquiring competence to legislate in the cybersecurity field: the Proposal for a NIS Directive [35] outlines the cascading effects across borders resulting from the intrinsic transnational dimension of NIS that a disruption of NIS may have and which affect the cross-border movement of goods, services and people. The 'disparities resulting from uneven NIS national capabilities, policies and level of protection across the Member States' are recognized as a barrier to the functioning of the internal market, and hence justifying EU action [35].

While in the internal market, the so-called first pillar, there is a rather broad legislative competence to regulate, this is not the case in the three other pillars, namely the Area of Freedom, Security and Justice (AFSJ), the Common Security and Defence Policy (CSDP) and the Common Foreign and Security Policy (CFSP)[9]. Legislation in the AFSJ is under Art. 83(1) TFEU mainly restricted to law enforcement [36], while in the CSDP the realization of a common cyber defence policy is presented with institutional challenges

8 —— Treaty on the Functioning of the European Union, OJ C 326, 26.10.2012, p. 47. Under Art. 114 TFEU, the EU can adopt 'measures for the approximation of the provisions laid down by law, regulation or administrative action in Member States which have as their object the establishment and functioning of the internal market'.

9 —— Addressing these three pillars in detail would go beyond the scope of this paper which focuses on the internal market

10 —— Regulation (EU) 2019/943 of the European Parliament and of the Council of 05.06.2019 on the internal market for electricity (recast) (Electricity Regulation), OJ L 158, 14.06.2019, p. 54.

11 —— Art. 59(2)(e) Electricity Regulation.

12 —— The EU Telecoms Framework consisted of Directive 2002/19/EC (Access Directive), Directive 2002/20/EC (Authorisation Directive), Directive 2002/21/EC (Framework Directive), Directive 2002/22/EC (Universal Service Directive), Directive 2002/58/EC (e-Privacy Directive).

13 —— The EU Telecoms Package consisted of Directive 2009/140/EC (Better Regulation Directive), which amended the Framework, Authorisation and Access Directive, Directive 2009/136/EC (Citizens' Rights Directive), which amended the Universal services and e-Privacy Directive and Regulation No 1211/2009 establishing the Body of European Regulators for Electronic Communications (BEREC).

and national sovereignty concerns [36]. The adoption of legislation based on the CFSP is legally excluded; accordingly, Council decisions are the most tangible instrument in this pillar.

2.3. Focus Internal Market: Sector-Specific Regulation

The afore outlined limited mandate of the EU to regulate cybersecurity resulted in a multitude of different European and national regulations as well as sector-specific standards that address different aspects of cybersecurity. Similar to general security aspects, cybersecurity was primarily addressed in sector-specific regulations that paid respect to sector specificities. For instance in the energy sector, the Electricity Regulation[10] requires the European Commission to develop a network code on cybersecurity of cross-border electricity flows[11].

Various sector-specific legislation introduced rules on the prevention and mitigation of security incidents. For instance, in the telecommunication sector, the EU Telecoms Framework[12], which had been amended by Telecoms Package[13], introduced rules on the prevention and mitigation of data breaches as well as notification obligations in the e-Privacy Directive[14], and added rules on security breaches in the Framework Directive[15]. The convergence of the telecommunications, media and information technology sectors resulted in the European Electronic Communications Code (EECC)[16] which covers all electronic communications networks and services by a single legal act and requires the implementation of technical (and organisational) security measures following a risk-based approach as well as the notification of security incidents of a certain quality[17].

In the financial sector, the Payment Services Directive 2 (PSD2)[18] introduced provisions on operational and security incidents affecting in particular electronic payments enabled by payment services providers.

Similar to the EECC in the telecoms sector, the DORA Proposal [37] aims to establish a single European legislation restricted to ICT and cybersecurity for all financial institutions by introducing a more harmonised and comprehensive framework that spells out requirements to address and mitigate ICT and cyber risks at the level of the financial sector.

Legislative action also targeted the (cyber)security of particular assets, as for instance mandatory security measures to ensure security of personal data[19].

14 — Directive 2002/58/EC of the European Parliament and of the Council of 12.07.2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (e-Privacy Directive), OJ L 201, 31.07.2002, p. 37. See Art. 4 e-Privacy Directive. Art. 4(2) provides for an obligation to inform the subscribers of a particular risk of a breach of the security of the network.

15 — Directive 2002/21/EC of the European Parliament and of the Council of 07.03.2002 on a common regulatory framework for electronic communications networks and services (Framework Directive), OJ L 108, 24.04.2002, p. 33..

16 — Directive (EU) 2018/1972 of the European Parliament and of the Council of 11.12.2018 establishing the European Electronic Communications Code (Recast), OJ L 321, 17.12.2018, p. 36.

17 — See Art. 40 EECC.

18 — Directive 2015/2366/EU of the Parliament and of the Council of 25.11.2015 on Payment Services in the Internal Market, amending Directives 2002/65/EC, 2009/110/EC and 2013/36/EU and Regulation (EU) No 1093/2010, and repealing Directive 2007/64/EC, OJ L 337, 23.12.2015, p. 35.

19 —See Art. 32 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27.04.2016 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 (GDPR), OJ L 119, 04.05.2016, p. 1

3. Horizontal Approach to Regulating Cyber Aspects: NIS Directive and CRA

3.1. A Cross-Sectoral Approach Addressing Cybersecurity

While the legal measures and initiatives outlined in section 2.3. constitute sector-specific regulation, the NIS Directives and the CRA Proposal reflect a new approach in regulating cyber aspects by introducing rules on the underlying ICT infrastructure, hardware and software.

The first horizontal instrument, i.e. a cross-sectoral instrument, to regulate cybersecurity at EU level is the NIS Directive which entered into force in August 2016. The NIS Directive will soon be replaced by a NIS 2.0 Directive for which a Proposal was published along the 2020 Cybersecurity Strategy highlighting again the expedited speed of cybersecurity regulation since the NIS 2.0 Proposal was published six months ahead of the completion of the original foreseen first periodic review of the NIS Directive. Political agreement on a new Directive was reached in May 2022. Work on a CRA also intensified recently following the announcement of such a legislative action in the Commission 2022 work programme; a CRA Proposal was published in September 2022.

Similar to the NIS Directive, the NIS 2.0 Directive and the CRA follow a horizontal approach that addresses the underlying technology. While introducing provisions to make digital products more secure, the CRA will complement the NIS 2.0 Directive by also addressing manufacturers of tangible and intangible digital products and ancillary services. Previously, regulation has been aimed primarily at operators of ICT with the NIS Directive imposing obligations upon operators of essential services and digital service providers, and, for instance, the GDPR demanding state of the art security mechanisms to protect personal data.

The following section outlines how the NIS 2.0 Directive and CRA seek to improve cybersecurity and how they complement each other.

3.2. From NIS 1.0 to NIS 2.0: Imposing Obligations Upon Certain Operators of NIS

The NIS Directive: The original NIS Directive is a cornerstone of EU policy on cybersecurity laying down the foundations for a EU cybersecurity framework. The Proposal rooted in the Communication [38] released by the Commission in 2009 on critical information infrastructure protection from large scale cyberattacks[20] and the NIS Directive became a concrete deliverable of 2013 the Cybersecurity Strategy[21]. The choice of the legal instrument of a 'directive' means that the NIS Directive is not directly applicable in the EU Member States but binds the Member States as to the results to be achieved. The Member States have to transpose the Directive into the national legal framework leaving them a margin for manoeuvre as to the form and means of implementation. In the case of the NIS Directive, ENISA was tasked to assist MS to implement the Directive and support the strengthening of cybersecurity capabilities at EU level[22].

The NIS Directive lays down measures with a view to achieve a high common level of NIS security within the Union so as to improve the functioning of the internal market. To that end, the Directive covers capacity building and planning requirements, exchange of information, cooperation and common security and incident notification requirements for operators of essential services (OESs) and digital service providers (DSPs). The Directive only applies to entities identified by Member States as OESs in the sectors energy, transport, banking, financial market infrastructures, health sector, drinking water supply and distribution, and digital infrastructure[23]; DSPs under the scope of the Directive are only those listed in Annex III to the

20 — In recognition of the economic and societal role of ICT infrastructures, the Communication noted that there is a clear need to rapidly put in place the necessary elements to build a framework that will feed into the future strategy for network and information security.

21 — A political agreement on the Directive was reached in 2015 after three years of negotiations between the co-legislators.

22 — In that regard, ENISA's mandate was further strengthened by the CSA.

23 — See Annex II.

Directive, namely, online marketplaces, online search engines, and cloud computing services.

As regards OESs, the Directive only requires a minimum level of harmonisation[24], recognising that the legal systems in some EU Member States had already set higher standards, or may aim for higher standards than those required by the Directive[25]. In contrast, as regards DSPs, the Directive employs a maximum harmonisation approach[26], meaning that Member States may not introduce rules that are stricter than those set in the Directive.

With the 2013 Cybersecurity calling for effective EU-wide co-operation, including between authorities, public and private sectors, the NIS Directive introduced several cooperation mechanisms. In particular, two new fora were created: the NIS Cooperation Group[27] (to support and facilitate the strategic cooperation and exchange of information among Member States) and a network of Computer Incident Response Teams[28] (CSIRTs) (to improve the handling of cross-border incidents, share information about risks and coordinate responses to specific incidents). Further, Member States are required to designate a single national central contact point (SPOC) as liaison office for supranational cooperation[29].

Article 7 NIS Directive also required Member States to adopt a national cybersecurity strategy defining the strategic objectives and appropriate policy and regulatory measures with a view to achieving and maintaining the desired high level of cybersecurity.

Deficits of NIS Directive: The review process of the NIS Directive that was conducted in 2021 identified limitations as well as deficiencies that are deemed to have prevented the NIS Directive from unlocking its initially foreseen full potential.

First of all, a weakness of NIS Directive is its limited scope of application, since the Directive only applies to certain DSPs and OESs; the latter restricted to the sectors energy, transport, banking, financial market infrastructures, health sector, drinking water supply and distribution, and digital infrastructure[30]. As a result, the NIS Directive fails to sufficiently address the increased interconnectedness and interdependencies in sectors outside its scope [39]. This potentially results in companies outside the scope of the Directive not sufficiently investing in cybersecurity because they are not legally obliged to fulfil a certain standard; however, the protection of these companies may be of similar importance, e.g. in the pharma industry or logistics, or because they are an important supplier of ICT to an OES or DSP [39]. A significant weakness of the Directive is also the broad discretion given to the Member States in defining the *de facto* scope of the Directive [39] as well as the vagueness of provisions and resulting unclear requirements [40]. This ultimately led to divergences across Member States in the implementation of the Directive and thus, a fragmented regulatory policy landscape. For instance, the leeway given to Member States in identifying OESs in the sectors encompassed by the NIS Directive resulted in national identification methodologies that differ significantly in terms of which types of services national authorities deem to be essential [40]. The way national thresholds are applied also varies across the EU [40][31]. As a consequence, similar entities are not treated consistently across the Union.

24 — See Art. 3 NIS Directive.

25 — In fact, various Member States have decided to include additional sectors (e.g. public administrations, postal sector, food sector, chemical and nuclear industry) and expand obligations for the sectors covered.

26 — See Art. 16(10) NIS Directive.

27 — The NIS Cooperation Group was established by Art. 11 NIS Directive with the aim to ensure strategic cooperation and information exchange among EU Member States.

28 — See Art. 12 NIS Directive. The national CSIRTs collaborate in the CSIRTs Network 'to contribute to developing confidence and trust between the Member States and to promote swift and effective operational cooperation'.

29 — Art. 8(3) NIS 1.0.

30 — See Annex II NIS Directive.

31 — Preliminary evidence from the review process also suggests that the divergence between Member States may be related to two factors: the delegation of the identification process to sectoral authorities (e.g. ministries, agencies) and the top-down versus bottom-up (self-identification) identification procedure.

Information sharing is a central element of the NIS Directive; however, in practice, the information sharing about incidents and vulnerabilities remains limited[32], although national competent authorities report improvements [41].

In order to increase the cyber resilience of OESs and DSPs, the NIS Directive foresees the implementation of security measures (following a risk-based approach) and introduces an obligation to report significant incidents[33]. As with the OES identification procedure, the transposition of the respective articles into national law varies significantly [39]. Without an obligation to ensure coherence with certification schemes[34], some Member States introduced detailed security requirements, while others provide no guidance at all.

Different approaches in the transposition and in some cases pre-existing legislation[35], are one reason why security measures and incident reporting requirements are currently inconsistent across Member States. Another reason is that there is no common set of criteria as to what is considered an appropriate security measure in view of the risk posed and what is considered an incident [42]. Adding to uncertainties for reporting entities is the fragmented supervisory landscape [39].

Also, the review process identified different approaches to enforcement, inter alia in terms of regime of sanctions and penalties [39][36].

Besides the magnitude of obligations imposed on Member States, an impact assessment [43] in 2020 identified inter alia a low level of cyber resilience of businesses operating in the EU as well as inconsistent resilience across Member States and sectors.

The NIS 2.0 Proposal[37]: The NIS 2.0 Directive replaces the existing NIS Directive. A key change of the Proposal relates to its scope with new sectors being added and the Directive abolishing the differentiation between OESs and DSPs by introducing the concept of essential entities (EEs) and important entities (IEs). EEs are entities that operate in the sectors and sub-sectors listed in Annex I[38] or are of a type listed in Article 2(2)(a) NIS 2.0 Proposal. IEs are entities that operate in the sectors and sub-sectors listed in Annex II[39]. The Proposal tremendously extends the scope of application of the Directive by adding new sectors (inter alia include waste water, public administration entities, space and chemicals manufacture), amending existing sectors and also by setting a size-threshold. Member States will no longer be required to carry out an identification process to determine which entities meet the criteria to qualify as relevant operators. In order to eliminate the wide divergences among Member States in that regard, and to ensure legal certainty for risk management requirements and reporting obligations, a uniform size-cap rule is introduced whereby all medium and large entities (as defined by Commission Recommendation 2003/361/EC[40]), that operate within the sectors or provide the services covered by the Directive, fall within its scope. The Proposal replicates the minimum harmonisation approach under the existing Directive and extends this to all types of service providers[41].

The Proposal also replicates the obligation for Member States to adopt a national cybersecurity strategy[42]. In contrast to the existing Directive, Article 7 NIS 2.0 Proposal (consolidated text of November 2022) not only concretises the issues to be addressed but also provides a list of

32 — Operational information sharing focused on cross-border incidents, whereas the need to share information on vulnerabilities across the Member States to ensure more robust risk management is hardly addressed.

33 — Cf. Arts. 14 and 16 NIS Directive.

34 — As for instance set out in the CSA.

35 — Such as for instance the German IT security Act (Gesetzüber das Bundesamt für Sicherheit in der Informationstechnik (BSI-Gesetz)), which has only recently been extended by the IT-Sicherheitsgesetz 2.0 (IT Security Act 2.0).

36 — For the significant variation of the penalty levels see [52].

37 — All references in the section relate to the consolidated text adopted by the European Parliament in November 2022 [22].

38 — Annex I lists as 'sectors of high criticality': energy, transport, banking, financial market infrastructures, health, drinking water, waste water, digital infrastructure, ICT-service management (B2), public administration entities excluding the judiciary, parliaments and central banks, and space.

39 — Annex II lists as so called 'other critical sectors': postal and courier services, waste management, manufacture, production and distribution of chemicals, food production, processing and distribution, manufacturing, digital providers, and research.

40 — Commission Recommendation 2003/361/EC of 06.05.2003 concerning the definition of micro, small and medium-sized enterprises, OJ L 124, 20.05.2003, p. 36.

41 — See Art. 5 NIS 2.0 Proposal (consolidated text of November 2022).

42 — Art. 7 NIS 2.0 Proposal (consolidated text of November 2022). This requirement was previously referred to as 'national strategy on the security of network and information systems' (Art. 7 NIS Directive).

policies that Member States will have to adopt including, inter alia, a policy addressing cybersecurity in the supply chain for ICT products and services used by EEs and IEs, and a policy on the management of vulnerabilities.

As regards vulnerabilities disclosure, the NIS 2.0 Directive establishes a framework for so called coordinated vulnerability disclosure, where designated CSIRTs act as trusted intermediaries and thereby facilitate the interaction between reporting entities and manufacturers or providers of ICT products and services[43]. The confidential reporting of a vulnerability will also be possible for any natural or legal person[44]. Further, a European vulnerability database is set-up to which all interested parties shall have access[45].

At national level, Member States are required to have a national cybersecurity crisis management framework in place, inter alia by designating national competent authorities responsible for the management of large-scale cybersecurity incidents and crises[46].

Similar to the status quo, Member States are required to designate one or more national competent cybersecurity authorities for the Directive's supervisory tasks and a national single point of contact (SPOC) to exercise a liaison function in cross-border cooperation. The requirement to designate at least one CSIRT remains[47]. In contrast to the NIS Directive, the NIS 2.0 Proposal sets out an extensive catalogue of tasks for CSIRTs for the performance of which sufficient resources have to be allocated to the CSIRTs[48].

As regards cooperation at national level, the operative part of the NIS Directive only addressed cooperation between competent NIS authorities, the SPOC and the CSIRT(s) of the same Member State. The NIS 2.0 Proposal also addresses cooperation between these actors and law enforcement authorities, data protection authorities and further authorities[49]. The same actors are now also addressed in terms of cooperation at EU level in direct response to the perceived limited cooperation in practice[50].

Furthermore the tasks of the existing NIS Cooperation Group and the CSIRTs network are extended[51]. In order to support the coordinated management of large-scale cybersecurity incidents and crises at operational level and to ensure the regular exchange of relevant information among Member States and Union institutions, bodies and agencies, the NIS 2.0 Proposal also establishes the European Cyber Crises Liaison Organisation Network (EU-CyCLONe)[52]. Ultimately, in the field of cooperation, the Union is mandated to conclude international agreements in accordance with Article 218 TFEU with third countries or international organisations to allow and organise their participation in some activities of the NIS cooperation fora[53].

As a further new mechanism in the field of cooperation, the Proposal establishes a voluntary peer-review system with a view to, inter alia, learn from shared experiences, and strengthen mutual trust[54].

Cybersecurity risk management and reporting obligations remain a central element of the Directive. The Proposal requires Member States to provide that management bodies of the entities encompassed approve and oversee the cybersecurity risk management measures taken by the respective entities and to follow specific cybersecurity-related training[55]. The management may be personally liable for non-compliance with these obligations[56]. In terms of cybersecurity risk management, similar

43 — Art. 12(1) NIS 2.0 Proposal (consolidated text of November 2022).

44 — Ibid.

45 — Art. 12(2) NIS 2.0 Proposal (consolidated text of November 2022).

46 — Art. 9 NIS 2.0 Proposal (consolidated text of November 2022).

47 — Art. 10 NIS 2.0 Proposal (consolidated text of November 2022).

48 — Art. 11 NIS 2.0 Proposal (consolidated text of November 2022).

49 — Art. 13(4) NIS 2.0 Proposal (consolidated text of November 2022).

50 — Cf. Recital 107, which states that in relation to serious criminal activities, it is desirable that the European Cybercrime Centre and ENISA facilitate coordination.

51 — Arts. 14 and 15 NIS 2.0 Proposal (consolidated text of November 2022).

52 — Art. 16 NIS 2.0 Proposal (consolidated text of November 2022)

53 — Art. 17 NIS 2.0 Proposal (consolidated text of November 2022).

54 — Art. 19 NIS 2.0 Proposal (consolidated text of November 2022).

55 — Art. 20 NIS 2.0 Proposal (consolidated text of November 2022).

56 — See ibid. Similarly, any natural person responsible for or acting as a representative of an EE on the basis of the power to represent it will be held liable for breach of their duties to ensure compliance with the obligations laid down in the Directive, see Art. 32(6) NIS 2.0 Proposal (consolidated text of November 2022.

to the NIS Directive, Member States are required to ensure that entities encompassed take appropriate and proportionate technical, operational and organisational measures to manage the cybersecurity risks posed to the security of NIS[57]. In addition, the entities will in the future also be required to prevent or minimise the impact of incidents on recipients of their services and on other services[58]. The measures shall be based on an 'all-hazards approach' and the minimum measures are now outlined in the Directive. These include, inter alia, supply chain security, human resources security and business continuity measures[59]. The entities encompassed will also have to notify the national competent authorities or the CSIRTs of any cybersecurity incident having a significant impact on the provision of the service they provide. In order to demonstrate compliance with certain security requirements, Member States may require entities to use ICT products, services and process that are certified under European cybersecurity certification schemes adopted pursuant to Article 49 CSA60. Member States shall also encourage the use of European or internationally accepted standards and specifications[61].

In terms of critical supply chains[62], the Proposal introduces a requirement for the NIS Cooperation Group to conduct coordinated sectoral supply chain security assessments for particular technologies mirroring the risk assessment foreseen for 5G networks by the Commission Recommendation on Cybersecurity of 5G networks (EU) 2019/534[63]. The assessment shall take into account both technical and, where relevant[64], non-technical factors including those applied to 5G networks[65].

For the purpose of contributing to the security, stability and resilience of the DNS, TLD registries and the entities providing domain name registration services for the TLD shall collect and maintain accurate and complete domain name registration data[66].

While the NIS Directive requires OESs and DSPs to report incidents which have resulted in actual harm, the Proposal expands the reporting obligation to incidents that have caused or are 'capable of causing severe operational disruption of the service or financial losses for the entity concerned', as well as incidents that have affected or are 'capable of affecting other natural or legal persons by causing considerable material or non-material damage'[67]. As regards the reporting procedure, the Proposal lays down a three-stage approach[68] in order to strike a balance between swift reporting that helps to mitigate a potential spread of an incident, and in-depth reporting that draws lessons from incidents and improves the future resilience of NIS[69]. Where entities become aware of an incident, they will have to submit an initial warning within 24 hours, followed by an initial notification within 72 hours updating the information and indicating an initial assessment of the incident; a final report has to be submitted not later than one month thereafter, or where the incident is still on-going, a progress report and a final report one month after the incident has been handled[70].

In terms of jurisdiction, EEs and IEs will be under the jurisdiction of the Member State where they are established[71]. Providers of public electronic communication networks and providers of publicly available electronic communications services are excluded from this general rule; these entities are deemed to fall under the jurisdiction of the Member

57 — Art. 21(1) NIS 2.0 Proposal (consolidated text of November 2022).

58 — Ibid.

59 — Art. 21(2) NIS 2.0 Proposal (consolidated text of November 2022).

60 — Art. 24 NIS 2.0 Proposal (consolidated text of November 2022).

61 —Art. 25 NIS 2.0 Proposal (consolidated text of November 2022).

62 — To identify the supply chains that should be subject to a coordinated risk assessment, the following criteria should be taken into account: (i) the extent to which EEs and IEs use and rely on specific critical ICT services, systems or products; (ii) the relevance of specific critical ICT services, systems or products for performing critical or sensitive functions, including the processing of personal data; (iii) the availability of alternative ICT services, systems or products; (iv) the resilience of the overall supply chain of ICT services, systems or products against disruptive events and (v) for emerging ICT services, systems or products, their potential future significance for the entities' activities (Recital 47 NIS 2.0 Proposal).

63 — Commission Recommendation (EU) 2019/534 of 26.03.2019, Cybersecurity of 5G networks, OJ L 88, 29.03.2019, p. 42.

64 — Art. 22 NIS 2.0 Proposal (consolidated text of November 2022). Already in 2021, the German legislator introduced a trustworthiness assessment of the manufacturer of critical components that mirrors the coordinated risk assessment for critical supply chains.

65 — See Recital 90 NIS 2.0 Proposal (consolidated text of November 2022).

66 — Art. 28 NIS 2.0 Proposal (consolidated text of November 2022). Furthermore, such entities are required to provide efficient access to domain registration data for legitimate access seekers.

67 — Art. 23(3) NIS 2.0 Proposal (consolidated text of November 2022).

68 — The original Commission Proposal foresaw a two-stage reporting process.

69 — Cf. Recital 101 NIS 2.0 Proposal (consolidated text of November 2022).

70 — Art. 23(4) NIS 2.0 Proposal (consolidated text of November 2022)

71 — Recital 113 and Art. 26(1) NIS 2.0 Proposal (consolidated text of November 2022).

State in which they provide their services[72]. For certain types of entities[73], jurisdiction is established at the place of their main establishment[74].

ince the review of the NIS Directive revealed a reluctance to share information on cybersecurity threats and incidents, the NIS 2.0 Proposal introduces a separate chapter on information sharing. Chapter VI provides a legal basis for the voluntary sharing of relevant cybersecurity information. First of all, Member States shall provide rules enabling entities to engage in cybersecurity-related information sharing within the framework of specific cybersecurity information-sharing arrangements[75]. In addition, Member States shall allow EEs and IEs to report, on a voluntary basis, cyber threats, near misses and relevant incidents that do not meet the reporting thresholds for mandatory reporting[76]. Furthermore, entities outside the scope of this Directive shall be able to report, on a voluntary basis, significant incidents, cyber threats, or near misses[77].

Although the NIS Directive required Member States to ensure that the competent authorities have the necessary powers and means to assess the compliance with the security and notification requirements, the supervision and enforcement regime of the NIS Directive has proven ineffective [21]. Accordingly, the NIS 2.0 Proposal seeks to strengthen supervisory powers via a minimum list of actions and means for competent authorities. The new means include, inter alia, on-site inspections and off-site supervision, and regular targeted security audits[78]. While EEs will be subject to a full ex-ante supervisory regime, a lighter, ex-post only, approach will apply to IEs, mirroring the so-called 'light-touch' approach applied to DSPs under the NIS Directive [44]. Member States must ensure that the competent authorities, where exercising their enforcement powers have certain powers including the power to issue warning and binding instructions as well as the power to impose administrative fines[79]. Besides the sanctioning regime with administrative fines of a maximum of at least EUR 10,000,000 or 2 % of the total worldwide annual turnover[80], the Proposal also establishes responsibilities and sanctions directed at natural persons exercising managerial functions[81].

In line with the new permanent, and moreover extended, mandate for ENISA under the CSA[82], the NIS 2.0 Proposal foresees additional action areas for ENISA. These include the development and maintenance of a European vulnerability database83, the provision of the secretariat of the EU-CyCLONe[84], a biennial report on the state of cybersecurity in the EU[85], the support in the organisation of Member State peer reviews[86], the collection of aggregated incident data from Member States and the provision of technical guidance for comparable information[87], as well as the creation and maintenance of a registry of entities providing certain cross-border services[88].

3.3. The CRA Proposal: Imposing Obligations Upon Manufacturers of Products with Digital Elements

The CRA Proposal supplements the CSA and aims to make digital products and ancillary services more secure. In order to achieve this aim, the CRA, similar to the NIS Directive, introduces horizontal cybersecurity rules. These

72 — Art. 26(1)(a) NIS 2.0 Proposal (consolidated text of November 2022).

73 — DNS service providers, TLD name registries, providers of domain name registration services, cloud computing service providers, data centre providers, managed service providers, managed security service providers, content delivery network providers, as well as certain digital providers, and public administration entities.

74 — Art. 26(1)(b) NIS 2.0 Proposal (consolidated text of November 2022). According to Art. 26(2) the main establishment is where the decisions related to the cybersecurity risk management measures are predominantly taken, or, if this cannot be determined, the place where cybersecurity operations are carried out.

75 — Art. 29 NIS 2.0 Proposal (consolidated text of November 2022).

76 — Art. 30(1)(a) NIS 2.0 Proposal (consolidated text of November 2022).

77 — Art. 30(1)(b) NIS 2.0 Proposal (consolidated text of November 2022).

78 — Cf. Art. 32(2) NIS 2.0 Proposal (consolidated text of November 2022) in relation to EEs, and Art. 33(2) in relation to IEs.

79 — Cf. Art. 32(4) and (5) NIS 2.0 Proposal (consolidated text of November 2022) in relation to EEs, and Art. 33(4) and (5) in relation to IEs.

80 — Art. 34(4) NIS 2.0 Proposal (consolidated text of November 2022), applying to EEs; IEs are subject to administrative fines of a maximum of at least EUR 7,000,000 or 1.4% of the total worldwide annual turnover of the undertaking, see Art. 34(5).

81 — Art. 32(5)(b) and (6) NIS 2.0 Proposal (consolidated text of November 2022) in relation to EEs, and Art. 33(5) in connection with Art. 32(6) in relation to IEs. This does not extend to criminal or civil liability (cf. Recital 128)

82 — Under the extended mandate ENISA is tasked to assist the Member States and the Commission in the implementation of the revised NIS Directive.

83 – 88 — Art. 12(2) NIS 2.0 Proposal (consolidated text of November 2022).

rules apply to industry stakeholders, namely manufacturers, importers and distributors of tangible and intangible products with digital elements. The European Commission [45] notes four specific objectives of the CRA: (1) to ensure that manufacturers improve the security of products with digital elements since the design and development phase and throughout the whole lifecycle; (2) to ensure a coherent cybersecurity framework, facilitating compliance for hardware and software producers; (3) to enhance the transparency of security properties of products with digital elements, and (4) to enable businesses and consumers to use products with digital elements securely. As a Regulation, the CRA will become directly applicable in the EU Member States on its entry into force.

The CRA follows the so-called 'New Legislative Framework' (NLF)[89], which aims to improve the internal market for goods by improving market surveillance and boosting the quality of conformity assessments. The NLF inter alia sets out requirements for accreditation of conformity assessment bodies, and the market surveillance of products. A central principle are high-level essential requirements in terms of health and safety that products have to meet before they can be placed on the Internal Market; these requirements are then detailed by harmonised technical standards drafted by European Standardisation Organisations[90].

The CRA will apply to products with digital elements (i.e. any software or hardware product and its remote data processing solutions) 'whose intended or reasonably foreseeable use includes a direct or indirect logical or physical data connection to a device or network'[91]. While in the initial call for evidence for an impact assessment [46], the terminology of 'digital products and ancillary services' was used, the notion of 'products with digital elements' indicates a commitment to an even broader regulation [47].

In line with the NLF requirements, several obligations need to be fulfilled before and whilst placing a product with digital elements on the market. For instance, manufacturers, importers and distributors need to ensure that the product with digital elements is accompanied with appropriate instructions and information in a language that is easy to understand in order to ensure a safe use by the user[92].

As regards further requirements, the Proposal distinguishes between two product categories; products with digital elements as the default category, and critical products with digital elements, which are subdivided into two classes. All products have to comply with the essential cybersecurity requirements laid down in section I of Annex I to the CRA Proposal. These requirements include 'security requirements relating to the properties of products with digital elements', such as the absence of any known exploitable vulnerabilities, a secure by default configuration, or the possibility to address vulnerabilities through security updates, and 'vulnerability handling requirements such as regular tests and reviews of the security of the product'[93]. Hence, the CRA will make security by design mandatory.

Products with digital elements that amount to critical products are enlisted in Annex III to the CRA Proposal. Generally speaking, a product is considered critical if the negative impact of the exploitation of potential cybersecurity vulnerabilities in the product can be severe due to, amongst

89 — The 'NLF' consists of Regulation (EC) No 765/2008 of the European Parliament and of the Council of 09.07.2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93, OJ L 218, 13.08.2008, p. 30; Decision 768/2008 of the European Parliament and of the Council of 09.07.2008 on a common framework for the marketing of products, and repealing Council Decision 93/465/EEC, OJ L 218, 13.08.2008, p. 82; and Regulation (EU) No 2019/1020 of the European Parliament and of the Council of 20.06.2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011, OJ L 169, 25.06.2019, p. 1.

90 — On the alignment of the CRA with the NLF see [47].

91 — Art. 2(1) CRA Proposal. Exceptions are listed in subsections (2) to (5) and mainly address situations where sectoral rules achieve the same level of protection as the one provided by the CRA. As regards the notion 'products with digital elements', the Commission departed in the Proposal from its prior terminology of 'digital products and ancillary services' used in the call for evidence for an impact assessment [46].

92 — Arts. 10(10) CRA Proposal (with regard to manufacturers), 13(2)(c) (with regard to importers), 14(2)(b) (with regard to distributors).

93 — Annex I to the CRA Proposal.

others, the cybersecurity-related functionality, or the intended use[94]. In particular, vulnerabilities in products with digital elements that have a cybersecurity-related functionality, such as secure elements, can lead to a propagation of security issues throughout the supply chain, rendering the product critical in the sense of the CRA[95]. As regards critical products, the Proposal further distinguishes between two different classes with class II representing a greater risk than class I. The Commission is empowered to adopt delegated acts supplementing the CRA to specify the product category definitions in class I and II.

In any case, before placing a product on the market, manufacturers must carry out appropriate conformity assessment procedures[96]. For the default category, manufacturers will have to self-assess conformity[97]. Products with digital elements that are certified or for which a statement of conformity has been issued under a cybersecurity scheme pursuant to the CSA and which has been identified by the Commission in an implementing act, shall be presumed compliant with Annex I[98]. The same applies to products with digital elements, which are in conformity with harmonised standards or parts thereof. Class I products must adhere to the application of a harmonised standard or certification scheme as set out in the CSA, or complete a third-party assessment to demonstrate compliance[99]. Class II products must always complete a third-party conformity assessment[100]. In line with the NLF, the Proposal sets out requirements for national authorities responsible for conformity assessment bodies[101].

If the compliance of the product has been demonstrated, manufacturers shall draw up an 'EU declaration of conformity' and state that the fulfilment of the applicable essential requirements has been demonstrated[102]. The EU declaration of conformity shall, inter alia, contain the elements specified in the relevant conformity assessment, and shall be continuously updated[103]. By drawing up the declaration of conformity, the manufacturer assumes responsibility for the product's compliance[104]. Further, the manufacturer can affix a CE marking to the product that indicates that it assumes responsibility for the conformity with all applicable requirements[105].

The CRA will also introduce reporting obligations of manufacturers similar to those for IEs and EEs under the NIS 2.0 Proposal: manufacturers shall, without undue delay and in any event within 24 hours of becoming aware of it, notify ENISA any actively exploited vulnerability contained in products with digital elements, as well as incidents having an impact on the security of those products[106]. In order to ensure that users can react quickly to incidents having an impact on the security of their products, manufacturers should also inform their users about any such incident and, where applicable, about any corrective measures to mitigate the impact of the incident, for example by publishing relevant information on their websites or, where the manufacturer is able to contact the users and where justified by the risks, by reaching out to the users directly[107].

Manufacturers also have to ensure that vulnerabilities of the product are handled effectively for the expected product lifetime or for a period of five years from the placing of the product on the market, whichever is shorter[108]. The Proposal also mandates that manufacturers are transparent on cybersecurity aspects that need to be made known to customers[109].

94 — Recital 25 CRA Proposal. As regards the intended use, the use in an in an industrial setting or in the context of an EE of the type referred to in Annex I to the NIS 2.0 Proposal renders a product critical since the severity of the impact of a cybersecurity incident may be more severe.

95 — Ibid.

96 — Arts. 20 and 24 CRA Proposal.

97 — Art. 24(1) CRA Proposal.

98 — Art. 24(2) and Recital 39 CRA Proposal.

99 — Ibid.

100 — Art. 24(3) CRA Proposal.

101 —Arts. 25 et seq. CRA Proposal.

102 — Art. 10 CRA Proposal.

103 — Art. 10(2) CRA Proposal.

104 — Art. 10(4) CRA Proposal.

105 — Art. 22 CRA Proposal.

106 — Recitals 19 and 35, Art. 11(1) and (2) CRA Proposal.

107 — Recital 35, Art. 11(4) CRA Proposal.

108 — Art. 10(6) CRA Proposal.

109 — Art. 10 CRA Proposal.

As regards importers, Article 13 CRA Proposal requires importers, inter alia, to ensure conformity of the manufacturer with the essential requirements set out in Annex I before placing a product on the market. When identifying a vulnerability in a product with digital elements, importers are obliged to inform the manufacturer without undue delay about that vulnerability[110]. Similar obligations apply to distributors[111].

In line with the principles of the NLF[112], a legal framework within which market surveillance can be carried out is drawn up113. National market surveillance authorities carry out the tasks enlisted in relation to that Member State.

Where the market surveillance authority of a Member State has sufficient reasons to consider that a product with digital elements, including its vulnerability handling, presents a significant cybersecurity risk, it shall carry out a product evaluation in respect of the product's compliance with the requirements laid down in the CRA[114]. Under certain circumstances, for instance, when there is a risk to the provision of the services by an EE, the European Commission may requires ENISA to carry out the evaluation[115]. The market surveillance authority has the power to impose or request the imposition of administrative fines. In that regard, the CRA Proposal establishes maximum levels for administrative fines that should be provided in national laws for non-compliance with the obligations introduced by the Regulation[116]. In order to ensure that the regulatory framework can be adapted where necessary, the power to adopt acts in accordance with Article 290 TFEU is delegated to the European Commission for, inter alia, updating the lists of critical products and specifying the definitions of these products, as well as specifying the minimum content of a EU declaration of conformity[117].

110 — Art. 13(6) CRA Proposal.

111 — See Art. 14 CRA Proposal.

112 — Here: Regulation (EU) No 2019/1020 of the European Parliament and of the Council of 20.06.2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011, OJ L 169, 25.06.2019, p. 1. See also [47].

113 — Arts. 41 et seq. CRA Proposal.

114 — Art. 43 CRA Proposal.

115 — Recital 59, Arts. 45 and 46 CRA Proposal.

116 —Art. 53 CRA Proposal.

117 — Arts. 50 et seq. CRA Proposal.

3.4. In Brief: The Interplay Between the NIS 2.0 Proposal and the CRA

The NIS 2.0 Proposal and the CRA Proposal both seek to regulate cybersecurity on a horizontal level with the CRA complementing the NIS 2.0 Directive in many aspects. Taking the example of supply chain security, the CRA recognises that cybersecurity of the entire supply chain can only be ensured if all its components are secure. Under the NIS 2.0 Directive, Member States have to address supply chain security in their national cybersecurity strategies, and ensure that supply chain security forms part of the mandatory security measures employed by EEs and IEs. Further, the Directive introduces a EU coordinated risk assessment of critical supply chains. However, due to the limited scope of application of the Directive, this leaves out a wide range of products with digital elements. In fact, most of the hardware and software products on the market are currently not covered by any EU legislation tackling their cybersecurity [45]. Accordingly, the CRA seeks to close the existing regulatory gaps: As regards the security of services provided by EEs and IEs the CRA will have a direct impact by facilitating the compliance with supply chain requirements in that the Regulation ensures that the products that EEs and IEs use for the provision

of their services are developed in a secure manner, and provided with security updates[118].

Correlation between the two legislative proposals will certainly arise in the field of vulnerability disclosure, incident reporting and information sharing. As outlined above, the CRA introduces reporting obligations similar to those for IEs and EEs under the NIS 2.0 Proposal. Once ENISA is made aware of an actively exploited vulnerability, it is requested to forward the notification to the relevant CSIRTs or, respectively, to the SPOCs designated under the NIS Directive as well as informing the relevant market surveillance authority[119]. Thereby ENISA ensures that the national CSIRTs and the SPOCs are provided with the information necessary to fulfil their tasks and raise the overall level of cybersecurity of EEs and Ies[120].

Where the information notified is relevant for the coordinated management of large-scale cybersecurity incidents and crises at an operational level, ENISA shall also submit the information to the EU-CYCLONe established by the NIS 2.0 Directive[121]. By this, the CRA supports the EU-CYCLONe in fulfilling its tasks under Article 14 NIS 2.0 Proposal. ENISA also prepares a biennial technical report on emerging trends regarding cybersecurity risks in products with digital elements for the NIS Cooperation Group[122], And thereby again contributes to the information gathering of a NIS cooperation forum. The CRA also encourages manufacturers of products with digital elements to consider disclosing fixed vulnerabilities to the European vulnerability database established under the NIS 2.0 Directive[123] – enriching the information sharing platform from which any natural and legal person may benefit.

118 — Cf. Recital 11 CRA Proposal.

119 — Art. 11(2) CRA Proposal.

120 — Cf. Recitals 19 and 34 CRA Proposal.

121 — Art. 11(3) CRA Proposal.

122 — Art. 10(6) CRA Proposal.

123 — Recital 34 CRA Proposal.

4. Conclusion/Outlook

While the EU's approach to cybersecurity has long being implemented in a piecemeal fashion, this approach has recently changed. Although the legislative landscape is still characterised by fragmentation and coexistence of national and European cybersecurity laws, the COVID-19 pandemic and further circumstances have triggered a change in how cybersecurity is addressed. Obviously this has been the result of the increased interconnectedness and interdependencies when it comes to technology.

Although already the NIS Directive introduced a horizontal approach to cybersecurity regulation, this did not prevent fragmentation across the EU. However, the systemic and structural changes introduced by the NIS 2.0 Directive amount to a fundamental shift of approach towards covering a wider segment of the economies across the EU. At the same time, the Proposal seeks to streamline the obligations imposed on the entities covered and to ensure a higher level of harmonisation responding to the fragmentation that the original NIS Directive resulted in.

The NIS 2.0 Proposal not only details an incident reporting procedure and strengthens the security requirements for the entities encompassed, it also entails measures aimed at improving policy building approaches at Member States level. New frameworks for supplier relationship risk management and coordinated vulnerability disclosure are introduced. While the NIS Directive aims at ensuring the continuity of services to guarantee

the proper functioning of the Union's economy and society, the building of cybersecurity capabilities across the EU and the mitigation of growing threats to NIS used by critical entities, it does not specifically address the cybersecurity of products. The cybersecurity of products is moreover an indirect consequence in terms of security of the supply chain. Similarly, the current EU regulatory framework on products, the NLF, does not address specifically the challenges linked to the cybersecurity of digital products. This existing regulatory gap will now be filled with the CRA.

Similar to the original NIS Directive in terms of services provided by OESs and DSPs, the CRA is the first EU-wide legislation that introduces common cybersecurity rules for manufacturers and developers of products with digital elements. Also similar to the NIS Directive in terms of the cybersecurity level of certain critical services, the CRA responds to a perceived low level of cybersecurity of products with digital elements throughout the product lifecycle. In that regard, it addresses the not only the product lifetime but also the whole supply chain from manufacturers to importers and distributors.

The outlined proposals are exemplary for an overall tendency to align legislation and reduce complexity between different, often sectoral regulatory approaches to cybersecurity. The common denominator for both proposals is that they address the underlying technology on a risk-based approach.

The described interplay between the initiatives shows an effort for a coherent approach to cybersecurity at EU level that closes regulatory security gaps in the digital value chain and eliminates conflicting or overlapping regulations [48]. At this point, it is worth highlighting again that compliance with cybersecurity requirements under the CRA does not stop once a product is placed on the market. While the EU legislation on products usually relies on the concept of 'placing a product on the market'124, the CRA explicitly implements a product lifecycle approach since technological products may evolve over time: they may become insecure, or may be applied in a new context. As regards the latter, it must be noted that technological advancement nowadays relates more to new technological application than progress in the basic underlying technology.

The Commission is optimistic about the CRA's potential to become an international point of reference beyond EU's internal market [49]. Clearly with its obligation upon manufactures and importers in terms of conformity with the cybersecurity requirements and the presumption of compliance when applying European certification schemes or standards, the CRA has the potential to push forward EU standards internationally and influence global markets. Also, the NIS Directive with its supply chain security elements and obligations imposed upon entities that offer essential or important services in the EU has extraterritorial reach. Legislation in the field of digital economy naturally influences global markets when drafted by an important market for data-driven businesses [50]. This externalisation of EU law has been referred to as the 'Brussels Effect' [50]. With the EU's market power in the digital economy, the data protection regime under the GDPR has proven a strong example of said effect due to both technical and economic non-divisibility of the products and services across global users.

Whether the CRA, or the NIS Directive will shape international standards in the same way as the GDPR remains to be seen [51, 52].

Ultimately, the legislative initiatives of the CRA and NIS 2.0 Directive have proven that although there is no explicit mandate for the EU to regulate cybersecurity, the existing legal basis allows for far-reaching horizontal legislation detached from sectoral security objectives. By following a risk-based approach, the instruments provide a regime that adapts the level of regulation to the risk level while at the same time providing a uniform approach to regulate the underlying technology. This approach is also reflected in the EECC and DORA Proposal (addressed in section 2) which also seek to harmonise a previously existing fragmented regulatory landscape.

Funding

The research for this article was funded by the Luxembourg National Research Fund (FNR) C18/IS/12639666/ EnCaViBS/Cole, https://www.fnr.lu/projects/the-eu-nis-directive-enhancing-cybersecurity-across-vital-business-sectors-encavibs/works.

References

[1] K. Okereafor, *Cybersecurity in the COVID-19 pandemic*. Boca Raton: CRC Press, 2021.

[2] Europol. (2021). *European Union serious and organised crime threat assessment 2021*. [Online]. Available: https://www.europol.europa.eu/cms/sites/default/files/documents/socta2021_1.pdf. [Accessed: Oct. 24, 2022].

[3] BBC. (2021, May 20). *Cyber-attack on Irish Health Service 'catastrophic'*. [Online]. Available: https://www.bbc.com/news/world-europe-57184977. [Accessed: Oct. 24 2022].

[4] BSI. (2021). *Die Lage der IT-Sicherheit in Deutschland 2021*. [Online]. Available: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2021.pdf%20?__blob=publicationFile&v=4. [Accessed: Oct. 24, 2022].

[5] BSI. (2022). *Die Lage der IT-Sicherheit in Deutschland 2022*. [Online]. Available: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2022.pdf?__blob=publicationFile&v=5. [Accessed: Oct. 24, 2022].

[6] European Parliament. (2021). *Recent cyber-attacks and the EU's Cybersecurity Strategy for the Digital Decade*. [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/ATAG/2021/690639/EPRS_ATA(2021)690639_EN.pdf. [Accessed: Oct. 24, 2022].

[7] Politiets Sikkerhetstjeneste. (2020, Dec. 8). *Datainnbruddet mot Stortinget er ferdig etterforsket*. [Online]. Available: https://www.pst.no/alle-artikler/pressemeldinger/datainnbruddet-mot-stortinget-er-ferdig-etterforsket/. [Accessed: Oct. 26, 2022].

[8] BBC. (2018, Feb. 28). *Fancy Bear: Germany investigates cyber-attack 'by Russians'*. [Online]. Available: https://www.bbc.com/news/world-middle-east-43232520. [Accessed: Oct. 26, 2022].

[9] L. Cerulus. (2021, Feb. 15). *France identifies Russia-linked hackers in large cyberattack*. [Online]. Available: https://www.politico.eu/article/france-cyber-agency-russia-attack-security-anssi/. [Accessed: Oct. 26, 2022].

[10] European Commission. (2022, Oct. 18). *Critical infrastructure: Commission accelerates work to build up European resilience*. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_6238. [Accessed: Oct. 24, 2022].

[11] J. Plucinska. (2022, Oct. 6). *Nord Stream Gas 'sabotage': Who's being blamed and why?* [Online].

Available: https://www.reuters.com/world/europe/qa-nord-stream-gas-sabotage-whos-being-blamed-why-2022-09-30/. [Accessed: Oct. 24, 2022].

[12] J. Thurau. (2022, Oct. 25). *Germany's critical infrastructure is poorly protected.* [Online]. Available: https://www.dw.com/en/germanys-critical-infrastructure-is-poorly-protected/a-63505983. [Accessed: Oct. 26, 2022].

[13] C. Vallance. (2022, May 10). *UK blames Russia for satellite Internet hack at start of war.* [Online]. Available: https://www.bbc.com/news/technology-61396331. [Accessed: Oct. 24, 2022].

[14] European Commission. (2022, Oct. 18). *Proposal for a Council Recommendation on a coordinated approach by the Union to strengthen the resilience of critical infrastructure, COM(2022) 551 final.* [Online]. Available: https://data.consilium.europa.eu/doc/document/ST-13713-2022-INIT/en/pdf. [Accessed: Oct. 24, 2022].

[15] European Commission. (2022, June 28). *Security Union: Commission welcomes today's political agreement on new rules to enhance the resilience of critical entities.* [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_4157. [Accessed: Oct. 26, 2022].

[16] European Commission. (2020, Dec. 16). *Proposal for a Directive of the European Parliament and of the Council on the resilience of critical entities, COM(2020) 829 final.* [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0829. [Accessed: Oct. 26, 2022].

[17] U. v. d. Leyen. (2021, Sep. 15). *2021 State of the Union Address.* [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_21_4701. [Accessed: Oct. 24, 2022].

[18] European Commission & High Representative of the European Union for Foreign Affairs and Security Policy. (2013). Joint Communication to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Cybersecruity Strategy of the European Union: An open, safe and secure cyberspace, JOIN(2013) 1 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=JOIN:2013:0001:FIN. [Accessed: Oct. 24, 2022].

[19] European Commission. (2015, May 6). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, A Digital Single MArket strategy for Europe, COM(2015) 192 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52015DC0192. [Accessed: Oct. 24, 2022].

[20] European Commission. (2022, May 13). *Commission welcomes political agreement on new rules on cybersecurity of network and information systems.* [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/IP_22_2985. [Accessed: Oct. 26, 2022].

[21] European Commission. (2020, Dec. 16). *Proposal for a Directive of the European Parliament and of the Council on measures for a high common level of cybersecurity across the Union, repealing Directive (EU) 2016/1148, COM(2020) 823 final.* [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A823%3AFIN. [Accessed: Oct. 26, 2022].

[22] European Parliament. (2022, Nov. 10). *Consolidated text and legislative resolution of 10 November 2022 on the proposal for a directive of the European Parliament and of the Council on measures for a high common level of cybersecurity across the Union, repealing Directive (EU) 2016/1148.* [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2022-0383_EN.html. [Accessed: Oct. 26, 2022].

[23] European Parliament. (2022). Legislative resolution of 10 November 2022 on the proposal for a regulation of the European Parliament and of the Council on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014 and (EU) No 909/2014 (COM(2020)0595 – C9-0304/2020 – 2020/0266(COD)). [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2022-0381_EN.pdf. [Accessed: Oct. 26, 2022].

[24] European Commission. (2022, Mar. 22). Proposal for a Regulation of the European Parliament and of the Council laying down measures on a high level of cybersecurity at the institutions, bodies, offices and agencies of the Union, COM(2022) 122 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0122. [Accessed: Oct. 26, 2022].

[25] European Commission. (2022, Mar. 22). Proposal for a Regulation of the European Parliament and of the Council on information security in the institutions, bodies, offices and agencies of the Union, COM(2022) 119 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0119. [Accessed: Oct. 26, 2022].

[26] European Commission. (2022, Sep. 15). Proposal for a Regulation of the European Parliament and of the Council on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/2010, COM(2022) 454 final. [Online]. Available: https://ec.europa.eu/newsroom/dae/redirection/document/89543. [Accessed: Oct. 11, 2022].

[27] G. G. Fuster, L. Jasmontaite, "Cybersecurity regulation in the European Union: The digital, the critical and fundamental rights," in *The Ethics of Cybersecurity*, M. Christen, B. Gordijn, M. Loi, Eds. Cham: Springer, 2020, pp. 97–115.

[28] European Commission & High Representative of the Union for Foreign Affairs and Security. (2017). Joint Communication to the Parliament and the Council, resilience, deterrence and defence: Building strong cybersecurity for the EU, JOIN(2017) 450. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017JC0450. [Accessed: Oct. 26, 2022].

[29] European Commission. (2019, Mar. 29). Commission Recommendation of 26 March 2019, cybersecurity of 5G networks, COM(2019) 2335 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019H0534. [Accessed: Oct. 26, 2022].

[30] A. Bendies. (2017). *A paradigm shift in the EU's common foreign and security policy*. [Online]. Available: https://www.swp-berlin.org/publications/products/research_papers/2017RP11_bdk.pdf. [Accessed: Oct. 26, 2022].

[31] European Commission. (2020, Feb. 19). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Shaping Europe's digital future, COM(2020) 67 final.* [Online]. Available: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52020DC0067. [Accessed: Oct. 26, 2022].

[32] European Commission. (2021, Aug. 24). *Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on the EU Security Union Strategy, COM(2020) 605 final.* [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021AE0879. [Accessed: Oct. 26, 2022].

[33] European Commission & High Representative of the Union for Foreign Affairs and Security Policy. (2020, Dec. 16). Joint Communication to the European Parliament and the Council, The EU's Cybersecurity Strategy for the Digital Decade, JOIN(2020) 18 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020JC0018&qid=1671800243772. [Accessed: Oct. 26, 2022].

[34] C. Calliess, A. Baumgarten, "Cybersecurity in the EU – the example of the financial sector: A legal perspective," *German Law Journal*, pp. 1149–1179, 2020, doi: 10.1017/glj.2020.67.

[35] European Commission. (2013, Feb. 7). *Proposal for a Directive of the European Parliament and of the Council concerning measures to ensure a high common level of network and information security across the Union, COM(2013) 048 final.* [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52013PC0048. [Accessed: Oct. 26, 2022].

[36] A. Bendiek, E. Pander Maat. (2019, Oct. 2). *The EU's regulatory approach to cybersecurity. [Online]. Available: https://www.swp-berlin.org/publications/products/arbeitspapiere/WP_Bendiek_Pander_Maat_EU_Approach_Cybersecurity.pdf.* [Accessed: Oct. 26, 2022].

[37] European Commission. (2020, Sep. 24). *Proposal for a Regulation of the European Parliament*

and of the Council on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014 and (EU) No 909/2014, COM(2020) 595 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0595. [Accessed: Oct. 26, 2022].

[38] European Commission. (2009, Mar. 30). *Communication from the Commission on Critical Information Infrastructure Protection "Protecting Europe from large-scale cyber-attacks and disruptions: enhancing preparedness, security and resilience", COM(2009) 149 final*. [Online]. Available: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0149:FIN:EN:PDF. [Accessed: Oct. 26, 2022].

[39] European Commission et al. (2021). *Study to support the review of Directive (EU) 2016/1148 concerning measures for a high common level of security of network and information systems across the Union (NIS Directive) – No 2020-665. Final study report. [*Online]. Available: https://www.ceps.eu/wp-content/uploads/2022/07/KK0921034ENN.en_compressed.pdf. [Accessed: Oct. 26, 2022].

[40] European Commission. (2021). *Report from the Commission to the European Parliament and to the Council assessing the consistency of the approaches taken by Member States in the identification of operators of essential services in accordance with Article 23(1) of Directive 2016/1148, COM(2019) 546 final*. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52019DC0546. [Accessed: Oct. 26, 2022].

[41] S. Schmitz-Berndt, A. Machalek. (2022). EnCaViBS – Summary report on cooperation. [Online]. Available: https://encavibs.uni.lu/wp-content/uploads/sites/158/2022/08/EnCaViBS-questionnaire-report_cooperation.pdf. [Accessed: Oct. 26, 2022].

[42] S. Schmitz-Berndt, "Defining the reporting threshold for a cybersecurity incident under the NIS Directive and the NIS 2 Directive", *Journal of Cybersecurity*, 2023 (forthcoming).

[43] European Commission. (2020, Dec. 16). Commission Staff working document, Impact assessment report, SWD(2020) 345 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020SC0345. [Accessed: Oct. 26, 2022].

[44] T. Sievers, "Proposal for a NIS Directive 2.0: Companies covered by the extended scope of application and their obligations," *International Cybersecurity Law Review*, vol. 2, pp. 223–231, 2021.

[45] European Commission. (2022, Sep. 15). *Cyber Resilience Act*. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act. [Accessed: Oct. 28, 2022].

[46] European Commission. (2022). "Call for evidence for an impact assessment," Ref. Ares, 1955751.

[47] P. G. Chiara, "The Cyber Resilience Act: The EU Commission's Proposal for a horizontal regulation on cybersecurity for products with digital elements," *International Cybersecurity Law Review*, pp. 255–272, 2022.

[48] bitkom. (2022). *Position paper on a Cyber Resilience Act (CRA)*. [Online]. Available: https://www.bitkom.org/sites/main/files/2022-05/20220519_CRA_Bitkom_Positionspapier_eng_final.pdf. [Accessed: Oct. 26, 2022].

[49] European Commission. (2022, Sep. 15). *State of the Union: EU Cyber Resilience Act – questions & answers*. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/QANDA_22_5375. [Accessed: Oct. 29, 2022].

[50] A. Bradford, „Digital economy," in *The Brussels Effect: How the European Union rules the world*. New York: Oxford University Press, 2020. pp.131–170.

[51] D. E. Sanger, N. Perlroth. (2021, June 8). *Pipeline attack yields urgent lessons about U.S. cybersecurity*. [Online]. Available: https://www.nytimes.com/2021/05/14/us/politics/pipeline-hack.html. [Accessed: Oct. 24, 2022].

[52] P. Contreras. (2022, June 8). *EnCaViBS poster series: NISD in a nutshell – penalties*. [Online]. Available: https://encavibs.uni.lu/2022/06/08/nisd-in-a-nutshell-penalties/. [Accessed: Oct. 26, 2022].

Cybersecurity is more than a Technological Matter – Towards Considering Critical Infrastructures as Socio-Technical Systems

Veronika Nowak | University of Vienna, SBA Research, Austria, ORCID: 0000-0003-4229-4134

Johanna Ullrich | University of Vienna, SBA Research, Austria, ORCID: 0000-0003-0297-9614

Edgar Weippl | University of Vienna, SBA Research, Austria, ORCID: 0000-0003-0665-6126

Abstract

Cybersecurity is still considered a purely technological challenge; however, despite all technological progress, this challenge remains unsolved – as emphasized by many high-impact attacks against public administration and industry worldwide. We postulate that the mere focus on technology fogs the bigger picture, since people generate, operate, and interact with all technological systems, thus making them socio-technical systems. Hence, in this commentary we argue for a change of perspective towards a holistic, interdisciplinary view on our technological infrastructure. By example of the European power grid – inarguably a critical infrastructure not only for daily life but also for the continuity of our polity – we show that through interpretation as a socio-technical system, systematic and interdisciplinary studies would allow to reveal how its (cyber)security is not only a technological matter. An interdisciplinary approach combining STEM disciplines and Social Sciences would additionally advance the understanding of stakeholders and their goals and mindsets as well as the manifold dependencies between technology and human actors. While interdisciplinary endeavours appear to be generally supported by funding agencies, reviewers, universities, and researchers, they rarely occur in practice. We discuss why this is the case and present ideas on how to facilitate more interdisciplinary research.

Corresponding author:
Veronika Nowak, University of Vienna, SBA Research, Austria;
ORCID: 0000-0003-4229-4134;
veronika.nowak@univie.ac.at

Keywords
critical infrastructure; cybersecurity; European power grid; socio-technical systems

Cite this article as: V. Nowak, J. Ullrich, E. Weippl, "Cybersecurity is more than a Technological Matter Towards Considering Critical Infrastructures as Socio-Technical Systems," ACIG, vol. 1, no. 1, pp. 169–175, 2022. DOI: 10.5604/01.3001.0016.2055

1. Introduction

Traditionally, researchers and practitioners in technological fields take a highly focused point of view on the systems they are working on. However, due to the now ubiquitous digitalization, everything becomes increasingly linked; therefore, isolated perspectives on individual infrastructures are no longer feasible. For instance, cyber-physical systems, i.e., legacy systems combined with digital infrastructure, are now prone to cyberattacks which threaten their reliable operation. This is also the case for power grids, a major critical infrastructure crucial for daily life as we know it.

Due to the many dependencies, today's energy supply faces additional challenges beyond ensuring cybersecurity: (I) The increased use of renewable energy makes power grids more volatile. Yet, the deployment of renewables is indispensable in responding to and fighting climate change; the drought of 2022, reducing power generation in Austrian hydroelectric stations [1] and French nuclear stations [2] alike, gave an impression of what to expect in the future. (II) Commodity markets influence producer structures and prices as the current gas shortage [3] – caused by Russia's full-scale invasion of Ukraine in February 2022 – shows. (III) Power supply does not only follow economic but also geopolitical rationales (see the emergency synchronization of Ukraine with the European grid in March 2022 [4] as well as the ongoing plans of the Baltics to become "electrically independent" from Russia [5]). (IV) Even a non-digitized power grid infrastructure would not be cyber-secure as more and more consumer devices become "smart", potentially causing high load fluctuations [6].

The secure operation of the power grid in presence of external factors like cyberthreats, climate change, economics and geopolitics requires multi-faceted expertise; therefore, we need specialists of diverse disciplines to work together. We argue that we do not only have to view critical infrastructures (like the power grid) as cyber-physical systems, but as socio-technical systems – thus considering the technological as well as the human factor and their mutual dependencies. To illustrate our point, we introduce a case study in which minor misfunctions and miscommunications cascaded and brought the European power grid on the verge of a blackout which would have left roughly 400 million people without electricity.

2. Critical Infrastructures as Socio-Technical Systems

In November 2006, the newly built cruise ship Norwegian Pearl was transported via the river Ems from its shipyard to the North Sea. As on

similar occasions before, a 380 kV power line crossing the river had to be disconnected. However, the disconnection was preponed on short notice; thereby, the responsible Transmission Grid Operator (TSO) failed to inform its neighbour TSOs whose lines had to serve as replacement. Consequently, a power line connecting the TSO with one of its neighbours suffered from high power flows. In response, the TSO decided to couple busbars of a nearby substation; unfortunately, this action did not have the desired effect, but further increased the line's power flow. Eventually, the line tripped automatically and started a cascade involving 32 tripped lines. This caused the European power grid to split into three unconnected islands. Two of them suffered from a lack of power generation because smaller generation units like wind turbines had automatically disconnected, which further complicated the situation. The third island experienced a surplus of generation because many smaller generation units had initially disconnected to facilitate stabilization and later automatically reconnected; moreover, the threat of further disintegration remained due to high power flows on certain lines [7]. The incident was eventually solved, but for 38 minutes 24 European countries stood at the abyss of a continent-wide blackout – "a national catastrophe" in which "after a few days [...] the population cannot be provided with necessary goods and services anymore" [8].

Two characteristics of our power grid infrastructure can be observed here: (I) initially minor incidents in everyday operations can easily multiply into cascading failures, and (II) for these failures, neither technology nor people are solely responsible. This brings to the fore yet another layer of our cyber-physical and, thus, interdisciplinary power grid systems: the human factor. We therefore propose a perspective which views our power grid as a socio-technical system. This means to go beyond the existing STEM interdisciplinarity and cooperate with the Social Sciences. For instance, Science and Technology Studies (STS) – a Social Sciences branch founded in the late 1970s/early 1980s – provides research into among others the evolvement of large technological systems [9] and how in such systems economic, technical, and social actors and – factors form a "seamless web" [10]. Regarding our contemporary, highly complex technologies, sociologist Charles Perrow pointed out that incidents are by no means unexpected, but "normal accidents" [11]. Failures are not due to human error alone, but also have their causes in the design and operation of such technologies. Organisational issues were also identified as causing one of the defining disasters in 20th century space flight: the explosion of the Challenger Space Shuttle in 1986 due to a malfunctioning technical component. Sociologist Diane Vaughan revealed in a detailed study that differences between and particularities of the company – and engineering cultures at NASA and the respective contractor led to the fatal launch decision [12]. Methods based on such a socio-technical perspective would be beneficial when analysing the everyday workings and (cyber)security requirements of critical infrastructure. Considering the power grid, an interdisciplinary approach jointly developed by STEM and STS researchers would study the grid infrastructure, the people who maintain and operate it, the policy makers who regulate everyday operation and incident response, and the consumers who depend on electricity.

In this light, the incident from November 2006 might be interpreted as a malfunction of a socio-technical system: The TSOs relied on divergent maximum values for electric current of the initially tripping line. This was despite the actual values having been previously exchanged in an official program and being mentioned in the coordination calls during the incident. For grid stabilization, interventions on the energy market are (according to German law) only allowed as an ultimate measure and were therefore considered too late. The dis/reconnection behaviour of small generation units was irrelevant until the policy-driven extension of renewables made them a significant part of generation and, thus, even able to hamper stabilization efforts.

This is just one example of how, e.g., engineering decisions not only concern technical issues, but have economic, geopolitical, and societal reasons and consequences. Thus, we need to shift our perspective and look at our critical infrastructure as socio-technical systems. To do this properly, we will have to combine expertise and methodologies from STEM disciplines and engineering with Social Sciences and Humanities (SSH).

3. Interdisciplinary research and its challenges

In general, the need for alternative perspectives and interdisciplinary approaches has been acknowledged. In Austria, this is also visible in the funding landscape: the KIRAS [13] funding scheme, managed by the Austrian Research Promotion Agency (FFG) and covering security research, requires a consortium partner from SSH. The Austrian Science Fund (FWF) recently opened the funding call *Emerging Fields* [14] which aims at researchers who are "prepared to depart from established approaches" and specifically invites interdisciplinary teams to apply. In 2020, the Vienna Science and Technology Fund (WWTF) awarded roughly € 3.6 million to interdisciplinary projects [15] working within the framework of Digital Humanism [16]. Being grant applicants ourselves, we, however, experienced some difficulties with interdisciplinary proposals. Especially longer-term projects with ambitious approaches combining Computer – and Social Sciences did not succeed. Although the reviewers generally welcomed these interdisciplinary ideas, we got the impression that they did not know how to assess them or found them too risky for funding. As far as we can gather from such experiences and conversations with colleagues from various disciplines, almost everybody wants interdisciplinary research, but has little to no idea on how to do it.

Even to receive funding for projects combining STEM disciplines can be challenging, as the example of our *NurZu!* project shows. Its general idea is to combine the *Energiemosaik* – a model based on national statistics to infer power consumption of Austrian communities developed by the University of Natural Resources and Life Sciences (BOKU) – with an open data-based power grid model to analyse the power grid's resilience regarding climate change, cyberattacks, and diverging consumption/generation behaviour. The idea was well-received by the required industry partners: small stakeholders in the energy sector saw potential to accelerate the power grid's transition towards renewables; however, the project was not

accepted upon first submission. Reviews considered the approach to be "standard"; in fact, we combine standard procedures which creates novel, interdisciplinary challenges. We explained this in more detail for the resubmission and eventually got funded. To what extent this acceptance might be due to Russia's invasion of Ukraine – emphasizing the need for changes in the energy sector – would be a question for a socio-technical study.

Our experience from 15+ years working on information security in science and practice has shown that purely technical solutions to security issues often do not (fully) work as intended. Digitalization, almost always marketed as the cure to technological and often also societal problems, has turned out to be a curse as well. Functionality, a fundamental tenet of engineering, is usually put before security. While this is already troubling in the context of web applications and business software, it becomes a huge societal issue with regards to critical infrastructures. We therefore postulate that a broader perspective will be crucial to better understand those socio-technical systems, securely operate them, and meet future challenges. Still, apart from the lack of adequate funding schemes, there are issues regarding interdisciplinarity:

1. Concepts, terminology, and methods: even between STEM disciplines, the differences can be huge; just finding a joint vocabulary usually requires considerable efforts.

2. Discipline-specific customs regarding publications: e.g., importance of conference papers in Computer Science vs publishing journal papers and books in Social Sciences and Humanities (SSH).

3. Career stage of the involved researchers: while early-stage researchers might have flexibility and be interested in an interdisciplinary path, mid-term researchers could be hindered in their career since they depend on publishing in specific venues.

4. Review system: apart from reviewers usually being specialised in one discipline, it has to be noted that the review system is overcapacity as it is.

A major consequence of these challenges is that interdisciplinary projects often happen in a "subordination/service mode" [17]; i.e., one of the disciplines is only playing a supporting role. The main question is how we create an academic environment conducive to equitable interdisciplinary research.

4. Conclusion

Although steps are taken towards interdisciplinary research, progress is slow and funding schemes bringing together STEM fields and Social Sciences are scarce. There is a lot of ideological support for such project ideas which does not translate into actual funding. This leaves very promising research potentials untapped, since there are SSH disciplines like Technology Assessment and STS which explicitly study technology, thus

providing a link to STEM fields. Since large critical infrastructures like power grids should be considered socio-technical systems – i.e., technical and human factors playing equally important roles – joint work of STEM – and SSH researchers should be systematically promoted and funded.

To overcome the challenges of different terminologies, publication customs, and career requirements – thus arriving at the "integrative-synthesis mode" [17] of interdisciplinary research in which the understandings of different disciplines truly work together – some major systemic changes are necessary:

1. Funding should be available to give interdisciplinary teams enough time to find a joint vocabulary and understand each other's methods.

2. Coordination between reviewers of different disciplines should be facilitated for interdisciplinary proposals; alternative review structures such as juries may be feasible.

3. Established conferences and journals should be more open for interdisciplinary work to promote rather than hinder an interdisciplinary researcher's career.

We are aware that the current academic environment is not amenable to such drastic changes, especially not in the short run. Still, at this point the need for holistic perspectives on contemporary socio-technical issues is more than obvious. Since being well-versed in multiple disciplines like Gottfried Wilhelm Leibniz or Marie Skłodowska-Curie is virtually impossible in today's academia, we need the specialists to team up, and we need a funding – as well as publication environment which facilitates and rewards such efforts.

References

[1] Austrian Power Grid (APG). (2022, Aug. 25). *Trockenheit reduziert Stromproduktion aus Wasserkraft um 31 Prozent*. [Online]. Available: https://www.apg.at/news-presse/austrian-power-grid-apg-trockenheit-reduziert-stromproduktion-aus-wasserkraft-um-31-prozent/. [Accessed: Oct. 27, 2022].

[2] S. Sijelmassi. (2022, Sep. 8). *Face au réchauffement climatique, l'avenir de l'énergie nucléaire est-il menacé?, TV5 Monde*. [Online]. Available: https://information.tv5monde.com/info/face-au-rechauffement-climatique-l-avenir-de-l-energie-nucleaire-est-il-menace-466906. [Accessed: Oct. 27, 2022].

[3] BBC. (2022, Sep. 29). Nord Stream 1: How Russia is cutting gas supplies to Europe. [Online]. Available: https://www.bbc.com/news/world-europe-60131520. [Accessed: Oct. 27, 2022].

[4] European Network of Transmission System Operators for Electricity (ENTSO-E). (2022, Mar. 16). *Continental Europe successful synchronisation with Ukraine and Moldova power systems*. [Online]. Available: https://www.entsoe.eu/news/2022/03/16/continental-europe-successful-synchronisation-with-ukraine-and-moldova-power-systems/. [Accessed: Oct. 27, 2022].

[5] European Climate Infrastructure and Environment Executive Agency. (2021, June 18). *Baltic Synchronisation Project: Works are on track*. [Online]. Available: https://cinea.ec.europa.eu/news-events/news/baltic-synchronisation-project-works-are-track-2021-06-18_en. [Accessed: Oct. 27, 2022].

[6] A. Dabrowski, J. Ullrich, E. R. Weippl, "Grid Shock: Coordinated Load-Changing Attacks on Power Grids: The Non-Smart Power Grid is Vulnerable to Cyber Attacks as Well," in *Proceedings of the 33rd Annual Computer Security Applications Conference*, Orlando, Florida, USA, 2017, pp. 303–314, doi: 10.1145/3134600.3134639.

[7] UCTE. (2007). Final Report – System Disturbance on 4 November 2006, *Union for the Co-ordination of Transmission of Electricity*. [Online]. Available: https://eepublicdownloads.entsoe.eu/clean-documents/pre2015/publications/ce/otherreports/Final-Report-20070130.pdf. [Accessed: Oct. 27, 2022].

[8] T. Petermann, H. Bradke, A. Lüllmann, M. Poetzsch, U. Riehm, Eds., *Was bei einem Blackout geschieht: Folgen eines langandauernden und großflächigen Stromausfalls*. Berlin: Edition sigma, 2013.

[9] T. Hughes, "The evolution of large technological systems," in *The Social construction of technological systems: New directions in the sociology and history of technology.* Cambridge, Mass: MIT Press, 1987, pp. 51–82.

[10] T. P. Hughes, "The seamless web: Technology, science, etcetera, etcetera," *Social Studies of Science*, vol. 16, no. 2, pp. 281–292, 1986, doi: 10.1177/030631278601600200 4.

[11] C. Perrow, *Normal accidents: Living with high risk technologies*. Princeton: Princeton University Press, 2011.

[12] D. Vaughan, *The Challenger launch decision: Risky technology, culture, and deviance at NASA.* Chicago: University of Chicago Press, 2007.

[13] *KIRAS Sicherheitsforschung, Sicherheitsklammer.* [Online]. Available: https://www.kiras.at/. [Accessed: Oct. 27, 2022].

[14] *FWF Programmes, Emerging Fields*. [Online]. Available: https://www.fwf.ac.at/en/research-funding/fwf-programmes/emerging-fields. [Accessed: Oct. 27, 2022].

[15] *WWTF, Digital Humanism 2020.* [Online]. Available: https://www.wwtf.at/funding/programmes/ict/index.php?lang=EN#ICT20. [Accessed: Oct. 27, 2022].

[16] *Dighum, Vienna Manifesto on Digital Humanism.* [Online]. Available: https://dighum.ec.tuwien.ac.at/dighum-manifesto/. [Accessed: Oct. 27, 2022].

[17] A. Barry, G. Born, Eds., *Interdisciplinarity: Reconfigurations of the social and natural sciences.* London, New York: Routledge, 2013.

The Tragedy of Smart Cities in Egypt. How the Smart City is Used towards Political and Social Ordering and Exclusion

Šárka Waisová | **Department of Politics and International Relations, Faculty of Arts, University of West Bohemia, Pilsen, Czech Republic**, ORCID: 0000-0001-7197-9269

Abstract

Smart cities (SCs) are a new and rising phenomenon emerging across the globe. The present article focuses on the possible impact of SCs on socio-political life and structure, and the organisation of the target society. Here, SCs are critically considered as the spaces where people live, work and vote. The aim of the present article is to discuss SCs, and the digital technologies used in SCs, as a possible instrument of social and political ordering and of social exclusion. Drawing on empirical evidence from Egypt, particularly Egypt's new capital, the article sketches out how the smart city has been used by political and military authorities to socially and politically order and engineer society as well as effectively exclude certain groups, mainly political opponents. Life in the new smart capital has a Janus face. On the one hand, inhabitants of the city have access to excellent services, modern infrastructure, first-class education and health care, and high-tech digital technologies which other Egyptians do not benefit from. On the other hand, these inhabitants are under permanent control and are prisoners of the system. Living segregated, with less freedom than any other Egyptian citizens, they are excluded from natural life in the country and cannot experience any organic development of society.

Keywords

Egypt, exclusion, political and social ordering, smart cities, technology as instrument of control

Cite this article as: S. Waisova, "The Tragedy of Smart Cities in Egypt. How the Smart City is Used towards Political and Social Ordering and Exclusion," ACIG, vol. 1, no. 1, pp. 176–187, 2022. DOI: 10.5604/01.3001.0016.0985

Corresponding author:
Šárka Waisová, Department of Politics and International Relations, Faculty of Philosophy, West Bohemian University, Pilsen, Jungmannova 1, Plzeň, Czech Republic; ORCID: 0000-0001-7197-9269; waisova@kap.zcu.cz

1. Introduction

Smart cities are a new and rising phenomenon emerging across the globe, from Europe to Asia. They are built on greenfield sites, or smart technologies[1] are installed in old urban centres. Smart cities are projects based on using modern advanced technologies and devices to improve human lives. There is no agreed-upon definition of a smart city, but it can be said that SCs combine data and digital technology to make better decisions and improve the quality of life. Most literature about SCs list six characteristics of urban smartness: smart economy, smart mobility, smart governance, smart environment, smart living and smart people [1, 2]. Smart city observers [2, 3] state that layers work together to make a smart city. First is the technology base, which includes a critical mass of connected smartphones and sensors. The second layer consists of specific applications which enable users to translate raw data into alerts, insight and action. The third layer is usage by cities, companies and the public. The rising number of SCs together with experience of using smart technologies and devices in urban spaces gives rise to several questions and dilemmas, it has become a policy issue as well as a topic for academic research.

The present article looks into the possible impact of SCs on socio-political life, structure and organisation of the target society. The present debate indicates that there are two main ways of approaching SCs and thinking about their existence: 1) technocratic, which considers SCs a technical-material conglomerate of modern smart technologies and high-tech devices and 2) social, which considers SCs a space of human existence and social, economic and political interactions taking place in an environment using smart technologies and devices. The first approach tends to reduce the SC issue into a debate on technological aspects such as what type of cloud computing would be better to use for smart urbanism. The second approach rather ignores technical and technological issues, and looks at how smart technologies and SCs may influence various aspects of social, political and economic life. To see SCs as a social space opens the door to thinking about SCs in several ways: SCs and urban planning, SCs as an instrument for environmental management and sustainable development, SCs as a way to improve people's security, transport and infrastructure, and as a method of socio-political control and ordering[2]. It is this final issue which is the subject of the present article. The rising number of smart cities and increasing experience with smart technologies in urban life raise several questions about the risk of urban governance driven by technology.

Here, smart cities are critically assessed as the spaces where people live, work and vote. My aim in the present article is to discuss SCs as a possible instrument of social and political ordering and as a tool for social exclusion. In other words, I will concentrate on the third above mentioned layer, i.e. how actors use technologies and SCs. Social and political ordering can be defined as the sum of those methods by which authorities try to control, regulate, enforce and encourage conformity to norms, both formally and informally, and influence human behaviour to maintain a given order or to set up a different one, which will also meet the authorities' needs and projections. Social exclusion is defined as the

1 —— Which include big data, the Internet of Things, artificial intelligence, mechanical algorithms and cloud computing, to name but a few.

2 —— For several other ideas about smart cities, digital solutions and the quality of human life see the McKinsey Report from 2018 'Smart Cities: Digital Solutions for a more Livable Future' (www.mckinsey.com/capabilities/operations/our-insights/smart-cities-digital-solutions-for-a-more-livable-future).

process in which individuals or groups are blocked from various rights, opportunities or resources (e.g. housing and infrastructure, employment, healthcare and democratic engagement) that are normally available to other members of society [4]. Social exclusion may be direct or indirect and can occur on different levels. Here I argue that the smart city can become an instrument to structure and organise society which may result in social exclusion and even effective segregation. In cases where SCs are built on greenfield sites as new projects unrelated to previous settlements, culture or socio-political developments, they offer a unique opportunity to engineer, manage, order and control society, and to scale up inequalities.

The socio-political risks are already known and have been debated. These risks include various methods for restricting people's freedom. Such restrictions stem from the proliferation of smart technologies, for example, the use of direct biometric identifiers and smart watching technologies, such as location trackers and cameras, for excessive surveillance which is then combined with the creation of a system of rewards and punishments [5–7]. A case in point is contemporary China, where phone-tracking devices are now everywhere, the police have been creating some of the largest DNA databases in the world, and the authorities have been building upon facial recognition technology to collect voice print from the general public. The Chinese government's goal is to design a system based on smart technologies and digitalisation to maximise what the state can find out about a persons' identity, activities and social connections. This could ultimately help the authorities maintain socio-political order and hence maintain their rule [8].

However, there are several other problems, dilemmas and issues connected to extensive use of digital technologies and building SCs, particularly on greenfield sites. In cases where a city does not evolve organically step-by-step but is planned as one organised project, where people and institutions are relocated in one moment or in several waves based on digital technologies, a smart city may become an instrument to engineer society, to manipulate and exclude particular groups or individuals, in other words it becomes an opportunity to control and order everybody. Smart cities are most often presented as better places to live on account of using smart technologies, devices and applications in an urban space. However, they may become instruments of segregation, exclusion and ordering when they are accessible only for some and are used for extensive surveillance and policing. For the Chinese government the installation of smart technologies and construction of smart cities is unequivocally connected to their aim of controlling and eventually punishing citizens. However, there are other cases in the world where intentionality is not clear or doesn't even exist. Nevertheless, smart cities and smart technologies and devices have become an instrument of social and political ordering and exclusion. Drawing on empirical evidence from Egypt, particularly Egypt's new capital, this article scrutinises how smartness may be used to socially and politically order society and has the capacity to exclude certain groups and implement segregation. Using Egypt as a case study was motivated by three facts: firstly, there are only a few studies on Egypt's SCs, particularly on the social impact of SC development [9, 10]; secondly, the country receives generous financial support from international financial institutions

to build smart cities; and thirdly, there are more than three dozen SCs projected to be constructed in Egypt (Table 1) with a target population of more than 15 million. This means that in 2030, when Egypt's population is estimated to reach 130 million people, every tenth inhabitant will live in a smart city.

The case study of Egypt is divided into two parts: firstly, an introduction to the history of smart cities and the socio-political context of smart city development in Egypt; and secondly, empirical evidence on smart cities in Egypt was collected. Data and information about the New Administrative Capital has been used in this paper. The New Administrative Capital (NAC) was chosen to study because it is the most developed smart city project in Egypt, and it is also seen as the model for future smart cities in this country. Data and information about Egypt's smart cities, and particularly about the NAC, were collected in desk research based on analysis of documents, academic articles, newspaper stories and social network posts, this information was then triangulated through several interviews with people familiar with Egypt's political and social situation. The conclusion discusses the findings, particularly how the NAC smart city project was originally conceived as using smart technologies, devices and applications in conjunction with political measures to order its inhabitants. It was found that the NAC project was not participatory and that it does not take into account the needs of the poor, the vulnerable and women. Furthermore, it excludes people without clearly expressed loyalty to the el-Sisi administration. The New Administrative Capital became an instrument of segregation. Based on detailed screening, people are either allowed or denied the right to work and live in the NAC. Segregation and exclusion have several effects: even though those living in the NAC have access to better services and modern infrastructure as well as first-class education or better paid jobs, they have less freedom than Egyptians living outside the city. People living in the NAC are excluded from normal life in the country, they do not have the possibility of experiencing social development. Those living outside the NAC have, in many respects, become second-class citizens and representatives of the 'old republic'.

2. Egypt and Smart Cities

Egypt has been witnessing a rapid population growth for several decades; while in 1960 Egypt had 26.6 million inhabitants by 2000 there were 69 million people in the country, and in 2021, 104[3]. Forecasts for 2050 indicate that the Egyptian population will reach 190 million if current growth continues [11]. This population growth has been a challenge for the country's housing situation and access to food as well as other goods and services. In particular, urban housing shortage has become a serious issue. During recent decades, the price of construction land in Egypt has increased, and consequently, informal housing has mushroomed. Estimates [12] indicate that the country will need more than 8 million additional housing units by 2030 to satisfy demand. Egypt's governments reacted by transforming housing policies, in 1977 the New City initiative was launched, which was later implemented by the New Urban Communities Authority, established

3 —— For population statistics and projections see Wo-rld Bank (data.worldbank.org/indicator/SP.POP.TO-TL?locations=EG)

in 1979. Since then, there have been 31 cities built across Egypt. The new cities are divided into groups, called generations, based on the time built, urban conception and technology used. The fourth period of mega-city projects started after the Arab Spring turmoil calmed down and a military coup installed General el-Sisi as president. Fourth generation cities were announced and confirmed by several documents and presidential speeches. They have been declared as an official goal in one of the government's strategic documents: Egypt Vision 2030[4].

It is important to stress that all this has been going on in the background of the administration of President Abdel Fattah el-Sisi. General el-Sisi first took power in a 2013 coup while serving as Egypt's defense minister and armed forces commander. In the 2018 elections, Sisi received 97 percent of votes and remained in the presidential office. In 2019 constitutional amendments were adopted which added two years to Sisi's term, extending it to 2024, at which point he will be allowed to seek an additional six-year term. Parliamentary elections since 2014 have not been regarded as free, and elections for local councils have not been held since 2008 [13]. The policy making process as well as public life have been dominated by the security apparatus since the 2013 coup [14]. President el-Sisi started a series of political and economic reforms, these reforms include construction of several mega projects (called 'projects of national importance')[5]. Fourth generation cities are part of them.

The Politics of Smart Cities

The aim to construct a fourth generation of cities was announced in 2018. As of October 2022, 37 new cities to be built on greenfield sites were in planning or under development (Tab. 1.), and 24 existing cities were undergoing smart city transformation where smart technologies will be implemented into the existing urban arrangement and socio-political life. In fourth generation cities technology is presented as the only viable option to resolve many social, environmental and economic issues. However, Egypt has no official definition of fourth generation cities or smart cities. There have been steps to draw up a code for Smart Cities for Egypt but to date no documents have been accepted. The biggest project is the new capital city of the country – the so called New Administrative Capital – which is planned to replace the old Cairo.

The main contractors to build smart cities and other infrastructure megaprojects are construction companies owned by the military [15][6]. The companies equipping the cities with smart technologies are mainly from other countries. Foreign companies include Chinese tech-giants such as Huawei, but also big Western companies such as Honeywell or Siemens. In the next section, the construction, management and operation of the New Administrative City will be scrutinised.

4 —— Egypt Vision 2030 (mped.gov.eg/EgyptVision-?lang=en)

5 —— For the complete list see The Arab Republic of Egypt Presidency (www.presidency.eg/en/المشاريع-القومية/)

6 —— The military owned companies are more competitive, based on the decision from 2016 that the military and other security institutions were given exemptions in a new value-added tax (VAT) law enacted as part of IMF-inspired reforms. The law states that the military does not have to pay VAT on goods, equipment, machinery, services or raw materials needed for the purposes of armament, defense and national security. Furthermore, in 2015, nearly 600 hotels, resorts and other properties owned by the military were exempted from real estate taxes. Military companies also receive an exemption from import tariffs and from income taxes. Cargoes sent to military companies do not have to be inspected.

Table 1. Fourth Generation Cities on the greenfield site announced in 2018 or later [16].

City	Situation	Target population
New Administrative Capital + Badr City	First phase of the construction completed, in December 2021 relocation of ministries started, including redeployment of officers. Housing for civil servants and lower-middle class employees.	6.5 million 650,000
New Alamein	Construction started, first flats and businesses opened.	2.5 million
New Aswan	First phase completed, second phase to be completed by 2025.	850,000
New Mansoura	First phase completed.	3 million
Sheikh Zayed Extension	First phase completed.	675,000
	Plans exist for	
6 October, 10 Ramadan, 15 May, Assiut, Behira, Beni Suef, Dakahlia, Giza, Luxor, Marsa Matrouh, Minya, New Akhmim, New Borg Arab, New Damietta, New Farfara, New Ismailia, New Nubaria, New Salhia, New Sphinx, New Obour City, North Sinai, Port Said, Rosetta, Sadat, Shrouk, Tiba, Toshka, Quena		

The Tragedy of the New Administrative Capital: How the City and Digital Technologies are Used for Social Exclusion and Socio-Political Ordering

The construction of the NAC began in 2015. The city is located around 45 kilometres east of Cairo. The owner and developer of the NAC is the New Administrative Capital Company (ACUD), co-owned by the armed forces (51 percent) and the New Urban Communities Authority (49 percent), which is part of the Ministry of Housing. The NAC will accommodate all government ministries, the presidential office, parliament, the headquarters of all domestic banks, diplomatic missions, eight fourth-generation domestic universities, six international universities, several hospitals, an international airport, research and development centres, a business hub, religious buildings and a new, very modern public transport network including high-speed rail. The NAC will also house the State Strategic Command Centre Complex which includes the Egyptian army headquarters and the Ministry of Defence. To accommodate all the civil servants, military officers and labourers a new neighbouring city – Badr City – is being constructed at the same time. Badr City, 47 km from Cairo and 7 km from the NAC, now has 160,000 inhabitants, but is planned to be home to 650,000[7]. The relocation of ministries and their civil servants started in December 2021 after several delays. When the relocation was started, President el-Sisi

7 —— For details see New Urban Communities Authority (newcities.gov.eg/english/New_Communities/badr/default.aspx).

declared the moment 'the birth of a new state' [17]. The second phase of construction was announced in September 2022. The first phase of construction works was paid for by Egypt based on a domestic budget and several loans from international financial institutions. For the second phase, foreign investors are being sought (there are already some Chinese and Emirati developers in the city) and more loans from international banks are being requested [18].

At first sight, the project seems perfect – modern, technologically advanced, offering housing opportunities for thousands and having the potential to guarantee sustainable development. However, the opening of the city and the first wave of relocating civil servants to the NAC presents number of issues: there are strong signs that the NAC is the government's instrument to order citizens, to exclude some social groups and set up a system of social segregation based on political loyalty. The empirical evidence indicates that the ordering and exclusion are achieved by a combination of five strategies:

- ownership of estates and facilities in the city
- pricing of housing combined with granting of permission to buy property in the city
- granting permission to work in the city
- a closed and controlled system of ticket sales for public transport
- management of life in the city including the system of communication with citizens

Here, supported by the evidence from the NAC, all five strategies will be introduced and analysed.

Ownership of the estates and facilities in the city

As mentioned above, the owner and developer of the NAC is the New Administrative Capital Company (ACUD), co-owned by the armed forces (51 percent) and by the Ministry of Housing. The government will pay ACUD to rent ministries and other official buildings in the new office district [18]. The main constructors of the city facilities are army-owned companies, the military is also responsible for security and public safety in the NAC. Based on a presidential decree from 2014 which entered into force in 2021, the Egyptian military has the authority to secure so-called 'public and vital facilities'. The armed forces are allowed to assist the police in protecting public and vital facilities without specifying any time limit. Crime against public and vital facilities and properties is subject to the jurisdiction of military courts [19] which means civilian cases are transferred to these courts. Properties in the NAC were proclaimed to be public and vital facilities. This means that anybody who lives, works or visits the NAC will be under military jurisdiction.

House pricing combined with the granting of permission to buy property in the city

From the beginning the NAC project has targeted the upper-middle class; however, when apartments in the NAC and Badr City went on the market, it

was clear that only a small circle of Egyptians would be able to buy or rent a property in either of the new cities. Even though Badr City was presented as housing for civil servants working in the NAC, and the NAC itself was planned to include housing for civil servants as well as social housing, the prices in both cities are higher than what Egyptian civil servants can afford. Based on the government's decision, civil servants may apply for a subsidy from employers in order to rent a flat. To buy a property the permission of ACUD is needed. Thus, permission to buy a property, the price of flats[8] and the system based on an employers' subsidy (which is not automatic as civil servants who relocate to the NAC or Badr City have to apply for it) have created a system which controls who will populate the city. Except civil servants, the NAC is planned to be populated by military personnel[9], wealthy elites from the Gulf, and high-income residents of governorates outside Cairo.

Granting permission to work in the city

Even once all ministries have relocated to the NAC, not all civil servants are expected to relocate as well. To work in the NAC civil servants need permission. The government's decision concerning who gets permission to be transferred to the new capital is a closely monitored security process. Until now, over 50,000 civil servants have been selected. After undergoing a series of security checks, those selected were put through a training programme[10]. The training focused on raising national awareness and digital skills. It was also announced that the government would maintain a database of trained civil servants regarding employee capability for possible future training[11]. The system of permissions to relocate and work in the NAC has the potential to divide families (when one family member receives permission and another is rejected) and handicaps certain groups, particularly women Based on local cultural and religious rules it is almost impossible for single women to relocate if their parents are not allowed or cannot move into the NAC or Badr City. Thus, women who want to work in the NAC have to commute every day.

A closed and controlled system of ticket sales for public transport

For those allowed to work in the NAC but not allowed to buy or rent a flat, the only possibility is to commute every day. However, the commute is controlled. The system of tickets for public transport is closed and the sale is controlled: employees get public transport tickets from their employers, these tickets only cover trips between the home station of the employee and the station where the employer is located. The whole system is digitalised; passengers have e-tickets and each station has a digital check point to screen the movement of people across the city. This system is completed by a sophisticated surveillance system. While a smart city is expected to have the potential to support participation, engagement and community, the controlled system of movement in the NAC effectively limits meetings and talking outside of the work place and thus is not conducive to the development of social life.

8 —— Social conditions remain difficult in Egypt, with around 30 percent of the population living below the national poverty line (according to the household survey results for October 2019 – March 2020). The World Bank, 2021, Egypt's Economic Update, April and October 2021. Available at: https://thedocs.worldbank.org/en/c/9dbe40280b581a94ff950a11cab42fb3-0280012021/original/4-mpo-sm21-egypt-egy-kcm2.pdf; https://www.worldbank.org/en/country/egypt/publication/economic-update-october-2021#:~:text=Unemployment%20declined%20to%207.3%25%20by,popu-lation%2C%20further%20hindering%20poverty%20reduction

9 —— Read more in Al Monitor (https://www.al-monitor.com/originals/2022/08/more-hurdles-could-delay-opening-egypts-new-administrative-capital#ixzz7hOlIN376) or Egypt Today (www.egypttoday.com/Article/3/53793/Pros-Cons-of-investing-in-New-Administrative-Capital)

10 —— For more see Egyptian Streets (egyptianstreets.com/2019/01/22/vacating-cairo-50000-public-sector-employees-to-be-transferred-to-egypts-new-capital-city-by-2020/), Ahram Online (english.ahram.org.eg/NewsContent/1/0/330603/Egypt/0/Egypts-state-employees-to-be-assessed-ahead-of-tra.aspx) or Asharq Al-Awsat (english.aawsat.com/home/article/2728911/egypt-govt-trains-employees-ahead-move-new-administrative-capital)

11 —— For more see Egypt Independent (egyptindependent.com/egypt-finishes-assessing-1042-state-servants-before-transfer-to-new-capital/)

Management of the city including the system of communication with citizens

The NAC has been declared the high-tech model for Egypt's future and a green sustainable city, the best in Egypt. There are plans to integrate people through smart applications: Wi-Fi is to be accessible in public areas, residents will be able to use smart cards and apps to unlock doors, make payments and report complaints and problems. Infrastructure is planned to include a data transmission network, e-gates, smart utility management and specialised data centres, to name but a few. The NAC has its own grid control centre and there will also be a National Energy Control Centre located as the strategic command headquarters for all stations across Egypt[12]. The administration and operation of the city is run from the centralised and integrated City Operation Centre and Commander Control Centre. The data and information for both centres on developments in the city are collected via more than 6,000 cameras linked wirelessly to a command centre, mobile phone trackers, digital check points and digital control gates in public transport stations, to name but a few. The centre also includes technology to run video analysis[13]. As declared by ACUD, the aim is to 'monitor crowd and traffic congestion, detect incidents of theft, and observe suspicious people or objects which trigger automated alarms in an emergency situation' [20]. There are also new solutions tested in the NAC. The state-owned Egypt Post is planning to issue a one-stop-shop e-card (a so called 'citizen single card') which is planned to be used for cashless payment, thee is also an ID programme, including digital signature and personal identification for various purposes (health system, etc.) [20]. Egypt's authentication system will use vein biometrics. The technology shines a near-infrared light on users' fingers and captures the image of the vein pattern using small cameras installed in the scanner. The image is then processed, compressed and stored in a system [21]. All the data generated by the surveillance system and gathered and evaluated by the control centres is owned by ACUD, which has no civilian control. When Colonel General Mohamed Ahmed Zaki Mohamed became Minister of Defence in 2018, democratic civilian control of the armed forces was terminated. In the NAC military and state apparatuses grow through. Currently, there is only a limited social city life in the NAC, furthermore, the NAC is controlled through digital technologies and applications. City inhabitants do not know what city management will do with personal and other data, so they will live under an omnipresent threat of the loss of privacy.

12 —The grid control centre for NASC will be delivered by Siemens. Siemens is also involved in several other strategic projects in Egypt, such as laying a communications cable through the Red Sea, creation of the Industry 4.0 Innovation Centre in the new capital´s Knowledge City and delivering low-voltage control systems for the new capital´s Iconic Tower. For more see Trade.gov (www.trade.gov/country-commercial-guides/egypt-electricity-and-renewable-energy).

13 — These digital surveillance systems, technologies and devices are delivered by international tech-giants such as Honeywell, Huawei, SAP, Orange, IBM, Fibre Misr System and Mastercard.

3. Concluding Debate: May Egypt's Smart Cities be Better Places for Living?

The New Administrative Capital was opened in December 2021. Currently, it is partly populated and has become Egypt's first functioning city of the fourth generation. Planned as the model for other smart cities in Egypt, it is worthy of evaluation. The interest here is to assess the smart city as a possible instrument of social and political ordering and exclusion.

Before debating the case of the NAC it has to be said that digital technologies themselves and the smart city as an urban concept are neutral on their own, only in the hands of people, in Egypt's case in the hands of political and military authorities, may smart cities and digital technologies and devices serve as ordering instruments and tools for exclusion. Digitalisation of Egypt and implementation of smart solutions, particularly in the NAC, go far beyond the construction of a smart city for sustainable development and improvement of the quality of life. In the hands of Egypt's political and military elite smart solutions effectively produce 'a state within a state, a society within society' – a space and a group with different rules, opportunities and a different political and social order than the rest of the country. In Egypt's case, the smart technologies of the NAC work as an instrument to engineer society to create a 'new state' [15, 17]. Even though life in the NAC is still in its infancy and the situation can still evolve, it seems from present information that the people responsible for the city prefer technologies and applications which have the potential for control, such as crowd management, predictive policing and smart surveillance, over those without it, such as apps for bike sharing, smart parking or waste collection, to name but a few.

In the NAC only those with permission can work in the city, and only those with permission and money can find a home there. Movement across the city is restricted, authorities have the equipment, technologies and interest to control it. The system of technological control has been effectively completed by legislation combined with the system of ownership. Military forces are not only the majority owner of city property, but the property itself is designated as 'public and vital', as such it comes under military rule and protection. In fact, those living and working in the NAC are segregated from other citizens of the country. However, this does not mean that the situation of the inhabitants of the NAC is better than that of other Egyptians. Life in the NAC has a Janus face: on the one hand, inhabitants of the city (all digitally literate and with smart devices) have access to better services, modern smart infrastructure, first-class education and health care, and high-tech digital technologies; on the other hand, they are prisoners of the system with less freedom than many other citizens of the country. The smart technologies are not used for civic participation, protection of civil liberties or crime prevention. They are rather used for data-driven policing and deep control. The inhabitants of the NAC have only a limited possibility of living authentically and experiencing a natural and organic development of society. The NAC has become an instrument of segregation, exclusion and a source of social, political and economic inequality.

For a long time urban planning has been discussed as a political practice [6], however, by implementing smart technologies into urban living the debate has got a new impetus, and now several new questions have been raised about the impact of technology on social life and human freedoms. Egypt's case indicates that the implementation of smart cities and smart technologies into urban planning necessitates not only research and debate concerning the positive effects of smart cities, but also of the societal risks which smart technologies may give rise to. The present article opens the debate about risks such as the 'terror of technology', misuse of cybernetic control, and the smart city as a disciplinary strategy [22–25].

Funding

This article is the outcome of a project supported by the Czech Science Foundation, Grant No. 19-09443S.

References

[1] A. Caragliu, C. Del Bo, P. Nijkamp, "Smart cities in Europe," *Journal of Urban Technology,* vol. 18, no. 2, pp. 65–82, 2011 doi: 10.1080/10630732.2011.601117.

[2] P. Lombardi, S. Giordano, H. Farouh, W. Yousef, "Modelling the smart city performance, Innovation," *The European Journal of Social Science Research*, vol. 25, no. 2, pp. 137–149, 2012, doi: 10.1080/13511610.2012.660325.

[3] J. Woetzel, J. Remes, B. Boland, K. Lv, S. Sinha, G. Strube, J. Means, J. Law, A. Cadena, V. von der Tann. (2018). *Smart Cities: Digital Solutions for a more Livable Future*. [Online]. Available: www.mckinsey.com/capabilities/operations/our-insights/smart-cities-digital-solutions-for-a-more-livable-future. [Accessed: Sep. 26, 2022].

[4] M. Daly, H. Silver, "Social Exclusion and Social Capital: A Comparison and Critique," *Theory and Society,* vol. 37, no. 6, pp. 537–566, 2008, doi: 10.1007/s11186-008-9062-4.

[5] D. Glowacka, R. Youngs, A. Pintea, E. Wolosik, "Digital technologies as a means of repression and social control," *Study for European Parliament*, 2021.

[6] A. Vanolo, "Smart mentality: The Smart City as Disciplinary Strategy," *Urban Studies*, vol. 51, no. 5, pp. 883–898, 2014, doi: 10.1016/j.giq.2016.06.004.

[7] R. Williams, "Whose Streets? Our Streets! 2020-21 "Smart City" Cautionary Trends and 10 Calls to Action to Protect and Promote Democracy," *Belfer Center for Science and International Affairs*, 2021.

[8] New York Times. (2022, June 21). *China surveillance investigation.* [Online]. Available: www.nytimes.com/2022/06/21/world/asia/china-surveillance-investigation.html. [Accessed: Sep. 26, 2022].

[9] M.A. Ali, "Smart city policy in developing countries: Case study of the new administrative capital in Egypt," *Journal of Public Affairs*, e2774, 2021, doi: 10.1002/pa.2774.

[10] S.Y. Tan and A. Taeihagh, "Smart city governance in developing countries: A systematic literature review," *Sustainability*, vol. 12, no. 3, p. 899, 2020, doi:1 0.3390/su12030899.

[11] A. Tawadros. (2021, Aug. 4). *Director of the Planning Ministry's demographic center. Egypt Today.* [Online]. Available: https://www.egypttoday.com/Article/1/106627/Egypt%E2%80%99s-population-to-rise-to-190-million-by-2050-if. [Accessed: Sep. 26, 2022].

[12] UNHABITAT. (2020). *UNHABITAT supports Egypt on standards for creating smart cities*. [Online]. Available: https://unhabitat.org/un-habitat-supports-egypt-on-standards-for-creating-smart-cities. [Accessed: Sep. 26, 2022].

[13] Freedom House, *Egypt. Freedom in the World,* 2022.

[14] BTI (Bertelsmann Transformation Index). (2022). *Country Report, Egypt.* [Online]. Available: https://bti-project.org/en/reports/country-dashboard/EGY. [Accessed: Sep. 26, 2022].

[15] Reuters. (2018, May 16). F*rom war room to boardroom. Military firms flourish in Sisi's Egypt.* [Online]. Available: www.reuters.com/investigates/special-report/egypt-economy-military/. [Accessed: Sep. 26, 2022].

[16] *New Urban Communities Authority. New Cities.* [Online]. Available: http://newcities.gov.eg/english/New_Communities/default.aspx. [Accessed: Sep. 26, 2022].

[17] Reuters. (2021). *Egypt prepares to start to move to new capital. [*Online]. Available: www.reuters.com/article/us-egypt-new-capital-idUSKBN2B91X3. [Accessed: Sep. 26, 2022]

[18] New York Times. (2022, Oct. 8). *A new capital worthy of the Pharaohs rises in Egypt, but at what price?* [Online]. Available: www.nytimes.com/2022/10/08/world/middleeast/egypt-new-administrative-capital.html. [Accessed: Sep. 26, 2022].

[19] Human Rights Watch. (2021). *Egypt: Unprecedented expansion of military courts*. [Online]. Available: www.hrw.org/news/2014/11/17/egypt-unprecedented-expansion-military-courts. [Accessed: Sep. 26, 2022].

[20] MCIT (Ministry of Communications and Information Technology). (2022). *Digital Egypt.* [Online]. Available: https://mcit.gov.eg/en/Digital_Egypt. [Accessed: Sep. 26, 2022].

[21] Daily News Egypt. (Feb. 2, 2021). *Efforts to link Egypt´s national ID system with finger-vein recognition technology.* [Online]. Available: https://dailynewsegypt.com/2021/02/02/efforts-to-link-egypts-national-id-systemwith-finger-vein-recognition-technology [Accessed: Sep. 26, 2022].

[22] Ahram Online. (2021). 2021 Y*earender: The new republic*. [Online]. Available: https://english.ahram.org.eg/News/448838.aspx. [Accessed: Sep. 26, 2022].

[23] Daily News Egypt. (2021, Nov. 2). *Egypt accelerating efforts to achieve a digital transformation become digital innovation hub.* [Online]. Available: dailynewsegypt.com/2021/11/02/egypt-accelerating-efforts-to-achieve-a-digital-transformation-become-digital-innovation-hub-ceo-siemens-middle-east/. [Accessed: Sep. 26, 2022].

[24] Daily News Egypt. (2022, June 26). *Egypt starts new phase of digital services based on electronic signature.* [Online]. Available: dailynewsegypt.com/2022/07/26/egypt-starts-new-phase-of-digital-services-based-on-electronic-signature-mcit/. [Accessed: Sep. 26, 2022].

[25] S. Kamel, "The potential impact of digital transformation on Egypt," *ERF Working Paper Series*, no. 1488, 2021.

Utilizing Object Capabilities to Improve Web Application Security

Michael Koppmann | **SBA Research Vienna, Austria**, ORCID: 0000-0001-5699-8226

Christian Kudera | **SBA Research Vienna, Austria**, ORCID: 0000-0003-1772-039X

Michael Pucher | **University of Vienna, Austria**, ORCID: 0000-0003-0123-0214

Georg Merzdovnik | **SBA Research Vienna, Austria**, ORCID: 0000-0002-9955-7284

Abstract

Nowadays, more and more applications are built with web technologies, such as HTML, CSS, and JavaScript, which are then executed in browsers. The web is utilized as an operating system independent application platform. With this change, authorization models change and no longer depend on operating system accounts and underlying access controls and file permissions. Instead, these accounts are now implemented in the applications themselves, including all of the protective measures and security controls that are required for this. Because of the inherent complexity, flaws in the authorization logic are among the most common security vulnerabilities in web applications. Most applications are built on the concept of the Access-Control List (ACL), a security model that decides who can access a given object. *Object Capabilities*, transferable rights to perform operations on specific objects, have been proposed as an alternative to ACLs, since they are not susceptible to certain attacks prevalent for ACLs. While their use has been investigated for various domains, such as smart contracts, they have not been widely applied for web applications. In this paper, we therefore present a general overview of the capability-based authorization model and adapt those approaches for use in web applications. Based on a prototype implementation, we show the ways in which *Object Capabilities* may enhance security, while also offering insights into existing pitfalls and problems in porting such models to the web domain.

Corresponding author:
Michael Koppmann, SBA Research Vienna, Austria;
ORCID: 0000-0001-5699-8226;
mkoppmann@sba-research.org

Keywords
Object Capabilities, secure design patterns, web security

Cite this article as: M. Koppmann, C. Kudera, M. Pucher, G. Merzdovnik, "Utilizing Object Capabilities to Improve Web Application Security," ACIG, vol. 1, no. 1, pp. 188–209, 2022. DOI: 10.5604/01.3001.0016.0823

1. Introduction

Exploiting security vulnerabilities for criminal activities has become a business that costs companies worldwide billions of U.S. dollars a year [1]. By 2026, the cybersecurity market size is forecast to grow to 345.4 billion USD [2]. Issues in how the security model of modern web applications are designed form part of this problem. *The Open Web Application Security Project* (OWASP) "Top 10" provides a regularly updated list of the most common problems in web applications, listing *Broken Access Control* as the Top problem in 2021 [3]. This is especially problematic in current web applications, which are commonly built around the concept of *Access-Control Lists* (ACLs). In such systems, authorization and performing a request are distinct actions, and the application itself has ambient authority and power, granted by the underlying operating system. Many techniques and patterns were developed to mitigate risks and security problems that are inherent to this kind of authorization scheme.

In this paper, we present an alternative approach to authorization based on the concept of *Object Capabilities*, which is not susceptible to attacks common in ACL-based systems. By using *Object Capabilities*, the *Principle of Least Privilege* (PoLP) [4] is built-in by design, mitigating the risk of the most common web attacks. A capability is described as a token of authority. It is a reference to an object, including a set of privileges or permissions. This token is transferable and unforgeable [5]. Together with other techniques, this concept can be used to implement complete authorization frameworks.

The aim of this paper is to provide a collection of capability-based security patterns for web applications, which prevent certain classes of vulnerabilities by design. A requirement for those patterns is to provide practical benefits for real-world web applications. Therefore, we focus on maintaining compatibility with the currently existing ecosystem of software products.

In particular, the main contributions of this paper are as follows:

- We provide an overview of capability-based designs in other application domains.
- We utilize these design patterns to illustrate how *object capabilities* can be utilized to improve web application security.
- With our capability-based reference implementation, *Eselsohr*, we illustrate the feasibility of the design patterns in practice.
- We make the source code of *Eselsohr* publicly available[1].
- We discuss changes to Browsers that are necessary for full utilization of the benefits of *Object Capabilities* in web applications.

1 —— https://github.com/mkoppmann/eselsohr

2. Background

A key concept in security is the concept of authorization and different approaches try to tackle this problem from different angles. Examples of this include identity-based authorization models revolving around permissions of specific users, or *object capabilities* granting the permissions itself, without necessarily involving identities.

2.1. Identity-based Authorization Models

The most common identity-based authorization model is the ACL. It's a list of permissions associated with an object. It specifies which subjects are granted access to it and which operations they are allowed to perform. One way to visualize this is as columns in an *Access Control Matrix* [6]. There are other identity-based authorization models, such as *Role-based Access Control* (RBAC) [7] and *Attribute-based Access Control* (ABAC) [8]. For our purposes their differences are not significant, so we will use ACLs as an umbrella term for identity-based access control.

2.2. Object Capabilities

A capability is a *transferable* and *unforgeable* token. It contains a reference to an object or a resource and the set of allowed actions that can be performed on it. Because the word "capability" is so overloaded with different meanings and because this concept overlaps with many object-oriented programming principles, nowadays, the term *Object Capabilities* (OCAP) is generally preferred. Nonetheless, capabilities are not tied to object-oriented programming and have been in use in various different contexts, first being mentioned in 1966 while discussing concurrent programming [9], in addition to being the basis for sophisticated implementations of operating systems [10, 11], hardware [5], kernels [12], file systems [13], and more. Because capabilities combine both designation and authority—meaning that they specify a particular resource and what access is allowed—whoever possesses a capability can exercise its authority. One of the key properties of capabilities is the ability to transfer them. *Unforgeability* is another important aspect of capabilities. It guarantees that they can only be accessed via (I) creation, (II) transfer, and (III) *endowment* [14]. Capability-based systems follow the *Principle of Least Authority*[2] (PoLA), as it follows from using *object capabilities*. This allows for collaboration between untrusted parties, as the potential damage that can be caused by a malicious actor is reduced to a minimum.

2 — We use authority and privilege interchangeably in this work, although subtle differences exist.

3. Related Work

In the context of web applications, ideas for capability-based security are often modelled upon already established standards. For example, Bearcaps [15, 16] and Bearer URL [17] both rely on URI schemes to represent

capability tokens. Bearcaps are URIs with two parameters: The access token and the web URL, while Bearer URLs rely on a similar syntax to HTTP Basic Authentication URLs. Compliant browsers should not reveal the token parameter through UI elements/JavaScript and compliant web servers should not log the tokens. While these URI schemes could be modelled through JavaScript and the HTTP Authorization header, extensive browser and server support is needed to protect the Bearer tokens against cross-site scripting (XSS) attacks. Instead of building on URI schemes, Macaroons [18] make use of cookies, extending them with caveats that attenuate or confine them and thus rendering them more suitable for authority delegation purposes. Caveats are nested and chained HMACs, used to append restrictions to the cookies, restricting usage and marking the need for additional authentication proofs. Each appended caveat consists of a list of predicates and a request is only allowed if all of these predicates are fulfilled.

The *Grant Negotiation and Authorization Protocol* (GNAP) [19], formerly known as OAuth 3.0, is an in-progress next generation protocol tackling the authorization problem from a different angle. While it is based on the experience of implementing OAuth 2.0 in practice, it is not compatible with previous OAuth standards. GNAP supports features like a built-in concept for identities and the ability to differentiate between running instances of the same app. GNAP is a protocol that can work with the *Authorization Capabilities for Linked Data* (ZCAP-LD) [20] data model. ZCAP-LD combines *object capabilities* and *Linked Data Proofs* [21] to allow delegating authority in a distributed network by chaining together capability documents. Each document can be further restricted by adding caveats to them to restrict their scope, their lifetime, or to revoke them later on.

While these projects work on introducing capability-based authorization models, they do not make direct use of OCAP themselves. The OCAP community continues to improve their security model steadily, working on projects beyond web applications and collaborating with committees to standardize their techniques. One of the more recent approaches is *Endo* [22], a sandboxed and OCAP-safe subset of JavaScript. It offers protection against malicious third-party dependencies by explicitly locking privileged features, e.g. network interaction, behind capabilities. Reviewable policies are used to restrict the authority of dependencies to a minimum. As part of the overall project, a new ECMAScript standard called *ShadowRealm* is currently proposed, which would dampen the impact of XSS attacks [23] by running user-provided input in an environment void of capabilities.

Other OCAP applications include *The Spritely Project* [24], which uses *object capabilities* to build a platform for federated social networks. The CapTP protocol, the foundational layer for Spritely, enables OCAP in distributed computing and supports different transport layer protocols, e.g. peer-to-peer applications through Tor [25]. In a similar manner, Cap'n Proto [26] is a serialization format and RPC framework, which includes OCAP-based security as one of its core principles. On a lower level, Fuchsia [27] is an operating system developed by Google, which prioritizes security as one of its design goals from the start. It makes use of OCAP, by enforcing all system calls to go through their defined virtual *Dynamic Shared Object*

(vDSO) interfaces, allowing more fine-grained access control through capability handles, when compared to typical system call ABIs.

4. Capability-Based Design Patterns

A capability in the context of programming languages is a reference to some piece of data. Although in object-oriented languages this is usually a reference to an object, it is not in fact limited to that. A capability can also be a reference to a primitive type, a function, a closure, or other data types, and can therefore be used with other programming paradigms, such as functional programming. In order to guarantee unforgeability, the programming language being used has to support "safe" pointers. References are pointers to a specific address in memory where data are stored. The language must protect these, such that it is not possible to, for example, cast an integer to a pointer (as is the case for the C language). Only creation, transfer, and endowment should give access to object references.

Transferability can be given by the fact that data, or references, can be passed as arguments to functions. In general, OCAP-based development can be achieved by omitting global scope, passing arguments, and utilizing lexical scoping [28].

```
function mkCounter() {
let count = 0;
return Object.freeze({
increment: function () {
return count++;
},
decrement: function () {
return count--;
}
});
}
counter = mkCounter();
entryGuard.giveCounter(counter.increment);
exitGuard.giveCounter(counter.decrement);
```

Listing 1: Counter example in JavaScript showing OCAP programming

Listing 1 demonstrates a "counter" example in JavaScript [29]. The function `mkCounter` contains a mutable variable called count and returns a JavaScript object that contains two functions: one to increment the count, and one to decrement it. `Object.freeze` creates an immutable version of the object. The count variable is only accessible within this function, a closure, and the only way to manipulate it is by calling one of the two provided functions, since they have access to the variable, being in the lexical scope. This is equivalent to a class in object-oriented languages where mkCounter is the constructor, count a private instance variable, and increment and decrement the public API. Lines 14–16 illustrate how to use this as a security mechanism. Imagine Alice has access to two other objects called "entryGuard" and "exitGuard". The entry guard should only have the power to count up, while the exit guard should only count down.

Alice creates a new counter object and passes only the increment function to the entry guard and only the decrement function to the exit guard. This is also a demonstration of PoLA, as both guard objects are only being given the authority they need to do their job. Similar to design patterns in object-oriented design, several patterns emerged for programming with *object capabilities*. We have collected some of these patterns and adopted them for use within the web security context. The following sections describe four of these patterns in detail.

4.1. Revoker

Listing 2 shows an example of the Revoker pattern [29]. When Alice passes an object reference to Bob, she has no means of forcing the reference to be returned. Early research work assumed that this limitation is a downside of *object capabilities* [30]. The Revoker pattern demonstrates how revocability can still be provided in an OCAP system. The `mkRevocable` function takes a function as an argument and returns an object with two functions: `wrapper` and `revoke`. The `wrapper` function can be passed to other objects, in this example to Bob, who can then proceed to interact with the wrapped function. If Alice later regrets that decision because Bob started to act strange, she can call the revoke function, which she has kept to herself. Then, `revoke` will set the function in the closure to `null`, rendering all further requests to the wrapper by Bob unusable.

```
function mkRevocable(fn) {
return Object.freeze({
wrapper: function (...args) {
return fn(...args);
},
revoke: function () { fn = null; }
});
}
revokableFoo = mkRevocable(foo);
bob.bar(revokableFoo.wrapper);
revokableFoo.revoke();
```

Listing 2: Revocation example in JavaScript

4.2. Membrane

Listing 3 shows an example of the *Membrane pattern* [31]. This can be used at the boundaries of the program architecture, where input and output with the real world has to be provided. Membranes then can be used to reduce the authority within the program by limiting the surface of available powerful capabilities. The function `mkReadOnlyFile` takes a file object as an argument and returns a new object that only exposes a subset of the original available functions.

```
function mkReadOnlyFile(file) {
return Object.freeze({
read: file.read,
getLength: file.getLength
});
}
```

Listing 3 Membrane example in JavaScript

4.3. Sealer

Listing 4 shows an example of the *Sealer* pattern [32, 33]. This can be used to securely transfer data between multiple objects without revealing the content to everyone involved. Similar to the Counter example, the `mkSealer` function uses sealed as its private state. The variable `sealed` is a WeakMap, a key/value store, where objects are keys; it is also not enumerable. It returns an object with two functions: seal and unseal. The seal function expects an argument called `data` and creates an empty object called box. The object identity of box is then used as key for `sealed` and `data` is used as value. The variable box is then returned to the caller. The function unseal takes a box as an argument and uses it to extract the value from the map.

Alice passes a sealer to Bob, which he can then use to send Alice a secret. Bob does not have a reference to Alice but Carol has. Since Bob has a reference to Carol, he sends her his box called `secretForAlice`. Carol, having no access to the unseal function, cannot see which secrets are being passed around and sends the box to Alice. Alice can access the secret by calling unseal with the provided object key.

```
function mkSealer() {
let sealed = new WeakMap();
return Object.freeze({
seal: function (data) {
const box = {};
sealed.set(box, data);
return box;
},
unseal: function (box) {
return sealed.get(box);
}
});
}
// Alice
sealer = mkSealer();
bob.foo(sealer.seal);
// Bob
secretForAlice = sealer.seal("Hunter2".);
carol.bar(secretForAlice);
// Carol
alice.baz(secretForAlice);
// Alice
secretFromBob = sealer.unseal(secretForAlice);
```

Listing 4: Sealer example in JavaScript

4.4. Limited Use

Listing 5 shows an example for the *Limited Use* pattern. This is a variation of the Revoker pattern, where we restrict the longevity of *object capabilities*. Instead of having an explicit revoke function, it has a built-in counter state that is reduced by one each time the wrapped function is called.

```
function mkLimitedUse(numOfInvocations, fn) {
let usages = numOfInvocations;
return Object.freeze({
use: function (...args) {
if (usages > 0) {
fn(...args);
usages--;
} else {
return;
}
}
});
}
```

Listing 5: Limited Use example in Javascript

These patterns also compose well together; to create a one-time use, read-only, revocable file capability, these constructs only have to be stacked: `mkLimitedUse` (1, `mkReadOnlyFile` (`revokableFile`). In the end, it all boils down to two guiding principles that allow a reduction of authority in program development and the risk of intentional or accidental bad behaviour:

No usage of global mutable state. This enforces that mutable data must be passed explicitly between objects, allowing controlling the flow of authority, thus, reducing the amount of potentially malicious actors in a system, who can manipulate a given piece of information.

Only controlled communication with the outside world. Interactions involving input and output should be wrapped at the edges of the program's architecture, providing a safe subset of possible functions. This allows that OCAP rules can be enforced within the program itself, while possibly dangerous code sections stay small and auditable.

These principles are sometimes easier, sometimes harder to follow, depending on the programming language and tooling in use.

5. Eselsohr

To illustrate the feasibility of our introduced capability-based design patterns in practice, we implemented Eselsohr, a bookmark manager where URLs can be stored in collections, which can then be shared with other people. Eselsohr supports the following features:

- **No Requirement for User Accounts.** Eselsohr does not require users to register before they can use the web application. Performing privileged actions is done by providing access tokens in URLs.

- **Support for Multiple Collections.** A single Eselsohr instance can handle multiple collections without any concept of an account. Access is granted by authorization, which does not require authentication.

- **Shareable Permissions.** Access to a collection is granted with URLs. New URLs with reduced permission sets can be created by users. Links can expire or be revoked by their owners.

- **Simple Embeddability.** Privileged Eselsohr actions can be integrated into other web applications. This is possible because the designation of a resource is coupled with the authority to perform the action.

We highlight the applicability of the design patterns across languages, by implementing Eselsohr in Haskell, a statically typed, immutable by default, purely functional programming language with lazy evaluation [34]. Although Haskell is a functional programming language and lacks the concept of an object, while also favouring immutability, many object capability patterns can still be applied with certain modifications. As object capability languages also try to achieve functional purity [35], Haskell's strictness on the separation between values and effects fulfils this goal. Eselsohr uses types in two ways as a kind of capability for achieving the principle of least authority: as access tokens within the runtime and to limit the possible effects it can have. The following sections describe how a capability-based approach is used in different layers of the application to provide features, improve maintainability, or enhance security. The goal is to reduce the ambient authority and stick to the principle of least authority.

5.1. Types as Capabilities

In short, external authorization systems like OAuth2 work like this:

1. A client wants to access a resource. They must prove that they are authorized to do so.

2. The client presents claims, such as their identity and the requested scope, to an authorization service.

3. This service performs the necessary authorization checks and returns a crypto-graphically signed token to the client.

4. The client presents this token to the resource service, which verifies the validity of the token before the service allows access to the requested data.

Types can be used to simulate this behaviour without using any cryptography but secured by the runtime of the language [36]. This can be achieved by using types with private constructors, which are functions that can create values of that specific type but are not exported outside of their

respective module. Other modules, therefore, cannot create values of that type directly but have to use the exported constructor function. Within this private constructor, all necessary authorization checks can be performed.

```
module Authorization
( AccessToken
, getData
, AccessArticle(..)
, DeleteArticle(..)
, accessArticleToken
, deleteArticleToken
) where
{- import required types and functions -}
-- | Constructor of the 'AccessToken'. type
newtype AccessToken a = a
-- | Function to access the wrapped value
getData :: AccessToken a -> a
getData (AccessToken data) = data
data AccessArticle = AccessArticle Id
data DeleteArticle = DeleteArticle Id
-- | Authorization function that maybe returns
-- the requested accesstoken or nothing, depending
-- if the checks succeed or not.
accessArticleToken :: Id -> User -> Maybe (AccessToken AccessArticle)
accessArticleToken articleId principal =
if {- perform authorization checks -}
then Just (AccessToken (AccessArticle articleId))
else Nothing
```

Listing 6: Example authorization module

A privileged function, such as a database-accessing one, would not then accept plain values as arguments but values that are members of such authorization types. These types can be unwrapped to access the required argument to perform the requested action. Values of such types work as a proof that the required authorization check has happened in the past as immutability guarantees that no change could have happened in between. Listing 6 shows such an authorization module, which handles access tokens. The type AccessToken wraps a generic type a, whose constructor is not exported from the module. The exported function `getData` can be used to unwrap the contained value. The types `AccessArticle` and `DeleteArticle` represent permissions to access or delete articles from the database respectively. The function `accessArticleToken` expects an article Id and a User as an argument and then performs authorization checks. If it succeeds, a value with the type Maybe (`AccessToken AccessArticle`) is returned. The same happens for deleteArticleToken with Maybe (`AccessToken DeleteArticle`). The caller of those functions can then decide how to continue, depending on the result.

```
module Database where
{- import required types and functions -}
getArticle :: AccessToken AccessArticle -> IO Article
getArticle token = do
let (AccessArticle articleId) = getData token
getArticleFromDB articleId
updateArticle :: AccessToken AccessArticle -> Article -> IO ()
updateArticle token updatedArticle = do
let (AccessArticle articleId) = getData token
updateArticleFromDB articleId updatedArticle
deleteArticle :: AccessToken DeleteArticle -> IO ()
deleteArticle token = do
let (DeleteArticle articleId) = getData token
deleteArticleFromDB articleId
```

Listing 7: Example database module

Listing 7 shows how a module with privileged functions can use these access tokens to guarantee that the caller performed an authorization check. The functions getArticle and updateArticle are expecting a value of type `AccessToken AccessArticle` instead of a plain Id. Therefore, it is not possible to call this function without a preceding call to `accessArticleToken`. In `deleteArticle` a different authorization check is enforced by requiring a different type of access token.

```
unauthorizedAccessToken :: a -> AccessToken a
unauthorizedAccessToken perm = AccessToken perm
```

Listing 8: Creation of access tokens without authorization

Some use cases require the unauthorized call of privileged functions, such as the initial fetching of the user value, as it is required for performing the authorization checks. Listing 8 shows a function which takes an argument with a generic type and converts it into the `AccessToken` type. The prefix unauthorized serves as a hint and can be detected during code reviews or by automatic tooling. This technique cannot merely be used for authorization but also for validation of external data. For example, instead of representing an email address as a `String`, a specialized `Email` type can be created. The constructor of that type guarantees that it follows a specific pattern, such as containing an @ symbol. We can further split this type up into two separate types, representing a verified email and an email that the user has yet to verify. Functions that expect a verified email have static guarantees that the verification step has been performed.

Further examples include the differentiation between SQL queries and data through the use of types, so that user input cannot be accidentally concatenated; thus preventing SQL injection attacks. Another example is the enforcement of proper output encoding in an HTML layout engine to prevent cross-site scripting attacks, by requiring that user input has to be converted to an HTML type.

Using types that enforce invariants and which represent concepts in the domain of the application is a pattern also known as “value object” in the realm of domain-driven design [37]. These types are implementations

of the Wrapper pattern. Working on primitive, built-in types like `String`, `Bool`, or Integer allows for too much latitude within the application. They are missing restrictions and a proper guidance for developers, which leads to a more fragile architecture. Utilizing the type system as a quasi-state machine, specifying the direction in which data has to flow through the application, enables the development of a more formal application design.

5.2. Types for Explicit Side-effects

The IO type gives us too much ambient authority. Adhering to PoLA means that we want to reduce the possible effects to a minimum. Object capability languages, such as Monte, only allow the import of IO-providing functions at the entry point of a module [38]. These functions are then passed along as arguments until they get called. This reduces the number of side-effects to the ones declared at the entry point. Eselsohr achieves a similar explicitness by using a custom data type that contains IO and Haskell's type class system [39]. Haskell's mechanism to generalize behaviour and patterns over multiple data types is called a type class. Examples of other type classes are `Eq`, for checking equality; Ord, for checking the ordering of elements; or `Num`, for numeric operations. For our purposes, type classes can be seen as similar to interfaces in object-oriented languages. Listing 9 shows the type class of `Eq`, which includes the equal (==) and not equal (/=) functions.

```
(==) :: a -> a -> Bool
(/=) :: a -> a -> Bool
```

Listing 9: Type class of Eq

Listing 10 shows the type signature of a polymorphic function called uniq, which, as for the Unix command-line tool, removes repeated adjacent lines in a list of a's. It works with any type a as long as the type has an implementation of the `Eq` type class. To reduce the number of possible side-effects in our program, and to simulate the approach taken by object capability languages, Eselsohr uses a custom data type and type classes to explicitly declare all possible side-effects per function.

```
uniq :: Eq a => [a] -> [a]
```

Listing 10: Type signature of a polymorphic function with Eq constraint

Functions that work with side-effects use—instead of running in IO directly—a generic type `m`. This type is then constrained by type classes that represent effects. The business logic of the application therefore remains polymorphic, and can be either completely pure or emit effects, depending on the data type that implements those type classes. This custom data type is called `App` in Eselsohr. Listing 11 shows such a function. Here, `createArticle` takes a Uri as an argument and returns an Article in the polymorphic type m. This type is constrained by having an implementation for the classes `MonadScraper` and `MonadTime`. The first class provides the function `scrapWebsite`, while the second class provides `currentTime`.

```
createArticle :: (MonadScraper m, MonadTime m) => Uri -> m Article
createArticle uri = do
aTitle <- scrapWebsite uri
aCreated <- currentTime
pure (Article aTitle uri Unread aCreated)
```

Listing 11: Example function showing effect type classes for a custom monad

Listing 12 shows the implementation of the `MonadTime` class for the `App` monad, which uses an IO function from the time package. `MonadIO` is another generalization and requires our App monad to have the ability to run IO actions. Only a thin, auditable layer of pure IO functions now wraps the logic of our application. Environment variables are parsed into a configuration data structure, folders in the file system are prepared, the `App` monad is created, and the web server is started, which executes our application logic for every incoming request. As type classes turn into dictionaries with functions as values at compile time, which are then passed along implicitly [39], we simulate the behaviour of object capability languages like Monte, where passing side-effecting functions happens explicitly.

```
import Data.Time (UTCTime, getCurrentTime)
class (Monad m) => MonadTime m where
currentTime :: m UTCTime
instance MonadTime App where
currentTime = currentTimeImpl
currentTimeImpl :: MonadIO m => m UTCTime
currentTimeImpl = liftIO getCurrentTime
```

Listing 12: Example effect class representing access to the time packagemodule

This is a form of the Sealer pattern. Only functions that were given a specific type constraint can access the IO functions contained in the App monad. Other functions could accept IO functions as arguments or return them, thus enabling them to pass these capabilities through the application. However, without the necessary type constraints they are not powerful enough to execute the capabilities. We achieve a clear separation between pure functions, which cannot cause any harm, and actions having effects, which should be the primary target when conducting security audits.

5.3. URLs as Capabilities

Eselsohr uses web-keys to transmit access tokens over HTTP, which was a compromise in the design process. The web-key technique [40] is an access control technique developed within the context of the Waterken web server, which in turn is written in an OCAP subset of Java. With web-keys, every capability is assigned a different 64-bit string, which is passed alongside URLs in a Base32-encoded form. It is used within Eselsohr as a means of passing capability-tokens without involving JavaScript, but special care is needed to prevent the token from leaking in the HTTP Referer header.

```
https://eselsohr.example.org/articles/example-title?acc=QMANQJKQ-
CLCDXJT5DI[...]VIPHU52S72SPX2CI2GU
```

Listing 13: Example URL to access a single article resource

Listing 13 shows how a single article resource can be accessed via a capability URL. The acc query parameter is a Base32 and binary encoded Haskell data type containing the ID of the referenced file and the capability. Eselsohr does not have a concept of users; article collections are stored in separated files and are identified by the access token. Alternatively, two separate query or path parameters could be used to avoid the need for serialization. The referenced capability has an optional expiration date, set to one month by default, and contains a reference to a resource, like an overview page or a single article, including a set of permissions. When an endpoint is called, types enforce the checking of these permissions before any resource can be loaded.

Users can create new URLs for each page with restricted permissions and expiration dates, implemented with a combination of the Membrane[3] and the Limited Use pattern. This makes it possible to create, for instance, time-limited read-only or append-only access to certain resources. Working with fine-grained permissions allows for dynamic use-cases that are hard to implement in static, group-based, coarse-grained scenarios. The generated URLs can also be revoked at any time, giving the owner full control over the access management by using the Revoker pattern.

3 —— A variation of the Membrane pattern had to be used, where a list of permissions is used, instead of embedding functions directly, as the capabilities had to be serializable for persistence.

The initial capability given to a person after creating a new collection is the entry point to the application. People are encouraged to store this URL somewhere safe, such as inside a password manager. This link to the initial overview page replaces the login page, since people do not have to authenticate themselves in order to use Eselsohr.

Omitting a central login page grants Eselsohr another property: its resistance to phishing attacks. By combining the designation and authority into a URL, transmitted over an encrypted and authenticated HTTPS tunnel, the user agent has the burden of verifying the authenticity of this connection. Traditionally, with usernames and passwords, the user is responsible for identifying whether the login form belongs to the right actor.

6. Evaluation

To reliably evaluate our proposed capability-based design patterns, we compare conceptual differences between Eselsohr and the open-source alternatives *Wallabag* [41] and *Espial* [42], both built on identity-based security paradigms. Wallabag was started in 2013 and is written in PHP. It is built on the Symfony [43] web framework. Wallabag can extract the content of web pages and display it in a more user-friendly format. Several import and export functions are available that aid migration to or from the service. Espial was started in 2019 and is written in Haskell and PureScript. The backend is built on the Yesod [44] web framework, while the frontend is written in PureScript [45], a Haskell-like language that transpiles to JavaScript and which is used to build *Single Page Applications* (SPAs).

Espial users can also add, in addition to web pages, notes with support for Markdown. Articles cannot only be added by browsing to the web application and using the corresponding form, but also by using a *bookmarklet*, a JavaScript snippet that can be bookmarked.

For the evaluation, we compare four main aspects of web applications: (i) user management, (ii) data manipulation, (iii) resource sharing capabilities, and (iv) embeddability and integration features.

6.1. User Management

All identity-based web applications require some kind of user or account model, which is then used for authentication and authorization. Additionally, secure password hashing algorithms, brute force protections, and session management are also needed. These common functionalities are often provided by the web framework in use. Both of the identity-based apps under consideration have built their user models on top of code provided by their chosen frameworks.

The OWASP list associates several risks with user management, including: (I) storing passwords in plain text, in encrypted form, or with a cryptographically weak hashing algorithm; (II) allowing weak or well-known passwords; (III) implementing a vulnerable password reset; (IV) missing brute force protections for the login procedure; (V) missing multi-factor authentication.

Because implementing secure user management is a non-trivial task, there is a trend in the industry of delegating this to third-party providers and using *Single Sign-On* (SSO) solutions like SAML or *OpenID Connect* for authentication and authorization [46]. Of course, this also increases the risks associated with centralization. If the same account, hosted by an identity provider, is used for a multitude of different services and access to it is removed temporarily or permanently, then either these applications can no longer be used or else the user has to start over with a new account.

With *object capabilities*, the concept of identity is optional. OCAP-based applications, like Eselsohr, do not require implementing user management and can therefore avoid the complexity and potential security issues associated with it.

6.2. Data Manipulation

Typically, identity-based web applications perform data manipulation with operations commonly known as “Create, Read, Update, Delete” (CRUD). For example, adding a new bookmark consists of:

1. The web application accepts a new request from the user.
2. A routing mechanism maps the URL path from the request to a controller, which handles requests for that particular path.
3. Inside the controller, data from the URL and HTTP body are optionally extracted if needed.

4. Some kind of authorization check is called to verify whether the calling user is allowed to perform the action.

5. Data are validated when handling user input, where it has to meet some criteria for further processing.

6. The controller calls a service performing the business logic or performs the logic itself; this typically involves a database.

7. Based on the return values of the service, a response is sent back to the user.

Checking if a subject is allowed to call a particular function or endpoint is called function-level authorization. Verifying that a subject is allowed to access a specific object is called object-level authorization. Validation of both of them for every single access is also called the Principle of *Complete Mediation* [47].

In Espial the function _handleFormSuccess[4] handles the main part of the controller's logic. It receives HTML form data from the user and starts by requiring that the current user has to have a valid authenticated session by calling requireAuthPair. This authorization function is provided by the used web framework. In this part of the application, object-level authorization is not used, as every user is implicitly allowed to add new bookmarks. The rest of the function then performs data validation, stores the new bookmark in the database, and archives the content of the bookmark. The main problem with this kind of controller logic is that the authorization logic is optional. There is no enforcement of access restriction to that endpoint or on objects themselves. In this case, the function requireAuthPair is also used for obtaining data about the currently logged-in user, so it is unlikely that this function be overlooked, yet there are other controllers where that could be the case. In addition, if Espial chooses to add a less privileged user group, which does not have permission to create new bookmarks, new authorization checks would then have to be added, without any guidance by the compiler or framework.

An OCAP-style web application, like Eselsohr, solves these problems on an architectural level. Access checks are statically enforced by embedding the result of authorization checks in type-level access tokens. Service code can then require such tokens, guaranteeing that authorization was successful. By using web-keys, it is not possible to designate a resource without the associated permission set, so we always fulfil the complete mediation principle.

4 —— https://github.com/jonschoning/espial/blob/c3a126b9eadb3c3778ab93ed4c4d0e80%35f669d3c/src/Handler/Add.hs#L59

6.3. Sharing Functionality

When using a web application, users have certain expectations compared to traditional desktop applications, such as the ability to share links to web pages or to bookmark them [40]. Identity-based applications usually only have the choice between public pages, which can be accessed by anyone,

or private pages, which require users to be logged-in when they open the link, since designation and authority are split.

This is the case in Espial. The bookmarks of users are public by default and are available at https://espial.example.org/u:username/. They also have the choice to declare a bookmark private, which hides it from that user's public page and requires an authenticated session. There is no functionality to share pages with a limited set of other people.

In Wallabag bookmarks are private by default, but they can be put into a public, read-only mode. In addition, *unread, archived, starred,* or *all* articles can be shared over RSS feeds. For this to work, Wallabag generates a 14-character-long random token as part of the feed's URL path, which acts as a capability, and which can also be revoked by the user at any time. It is not possible to generate multiple tokens or to place further restrictions upon them. The same token is also used for all available feeds, but it is still a very basic capability system, embedded in an otherwise identity-based application.

Web applications built on *object capabilities*, and web-keys specifically, take this concept further and allow everything to be shared with links if desired. For example, Eselsohr allows applying further restrictions on web-keys, such as a limited validity period and restricted permission sets.

6.4. Embeddability

The inability to embed identity-based products is also a weak point of them. HTML provides the functionality to embed other web pages with the iframe element, but its usefulness is often diminished for fear of security vulnerabilities. These frames are the main attack vector for clickjacking attacks and the main prevention method is by disabling the option to embed a web site completely, or at least for cross-origin requests. Once again, this works because an attacker can link to a well-known endpoint from a popular web site on their attack page and trick other people to reveal sensitive information or perform authorized actions, because the ambient authority granted by their browsers cannot differentiate between benevolent or malicious intent. This hinders the ability to build collaborative web applications.

Object capabilities allow for secure embedded pages and collaboration. Clickjacking, a confused deputy attack, is no risk for OCAP-based web applications, as an attacker would need to know the web-key for the page they want to link. At this point, they would already have access and have no need for social engineering techniques. Such applications can safely omit the HTTP headers that disallow framing the web site and allow other web applications to embed them as they like.

As an example, Eselsohr's new article functionality could be added as a custom widget in the instant messenger Element [48]. This messenger enables the embedding of arbitrary web pages as iframes by providing a link to them. Usually, only pages that do not require authorization can be used for this, as cookies cannot be used to show everyone in the channel the same page. However, web-keys do not have this limitation and such scenarios are therefore permitted. The provided web-key only has the

permission to create new articles for that specific resource and nothing else. If the link leaks, no other actions could be performed with it.

6.5. Discussion

The above comparison between Eselsohr and two identity-based projects shows that security vulnerabilities in authorization systems can be prevented with capabilities by design. By using types as authorization tokens, we obtain strong guarantees that authorization checks will not be overlooked and that services cannot be called unauthorized. Web-keys, a combination of designating a resource and a corresponding set of permissions, are not susceptible to confused deputy attacks and are therefore resistant to vulnerabilities like cross-site request forgery or clickjacking. Using the type system for explicit side-effects enhances the reasoning about the code base, since it provides a better understanding of which functions are safe to call and which are potentially dangerous. In addition, by disallowing arbitrary side-effects everywhere, certain areas of the program, such as those that handle untrusted user input with deserialization, become secure. In identity-based systems, subjects cannot choose which authority they want to use when accessing a resource. Authority is implicitly available in the environment and is granted based on the identity of the caller. In the context of code, this means that every function can potentially perform any action, as all code possesses equal authority. By requiring access tokens within the code, and thus making authority explicit, the authorization flow in the program becomes equivalent to the creation and passing of access tokens as arguments. Functions have to explicitly request authority before they can be used.

As subjects in identity-based systems do not have explicit control over authority, they cannot declare a purpose when accessing a resource. Therefore, a subject cannot securely perform actions on behalf of others, as all actions will use the authority of the subject. *Object capabilities*, on the other hand, combine designation with authority, thus allowing for the use case stated above. Collaboration is secure, since the subject is able to use each capability for its intended purpose. To uphold the principle of least authority, we want to grant subjects the minimum required amount of authority they need to perform their tasks. This can be done in an identity-based system by creating small identities with minimum rights, though it is hardly practical.

7. Conclusion

In this work, we proposed an alternative authorization model for web applications, utilizing *object capabilities*: object references combined with an associated set of permissions. We showed which security vulnerabilities arise when designation and authority are split apart in the context of web applications, and how this problem is inherent to applications built on access-control lists. In the addressed scenarios, we were then able to demonstrate that programming in an object capability style helps to

eliminate certain security vulnerability classes on an architectural level. Alongside this, we provided some techniques and patterns based on this style, such as web-keys.

A functional prototype was implemented to demonstrate that these techniques could be used in practice. The security analysis and model evaluation showed that OCAP-based applications have no significant drawbacks when compared to ACL-based applications, while providing improvements in areas such as shareability and embeddability. This was done by conducting a security evaluation, with a focus on the most common vulnerabilities found in web applications, and by comparing the prototype with other existing applications. We also looked at how modern browsers can securely exchange data between server and client, and which extensions are needed to better integrate and protect capabilities in web applications. The biggest problem remains that of transferring capabilities over URLs. Although hyperlinks are the primary method of navigating between web applications, web browsers currently assume that URLs only contain non-sensitive information, making it hard to embed sensitive information such as capabilities. In addition, a capability in a URL is a plain string with no further protection, so anyone could come up with a potentially valid capability, even though it was not passed to them explicitly.

The prototype was developed in a programming language that was not explicitly designed for this style of programming. It is capable of being run on common operating systems without the need for specialized application frameworks. As these existing systems are not following OCAP principles and assume ambient authority, adapters and wrappers are required to integrate them into an object capability application.

Finally, an overview of current OCAP-related projects was given, together with recommendations on how the prototype could be further improved in the future.

7.1. Future Work

To overcome the existing drawbacks, future research should evaluate how existing features in web browsers could be used to circumvent the current limitations in regard to transferring capabilities. Eselsohr made use of web-keys—capabilities in URLs—to navigate between web pages, because they work without JavaScript and can be used across different web applications. With JavaScript more methods of transfer would be available, such as adding HTTP headers, non-HTML HTTP bodies, or WebSockets. These channels could then be used to securely transfer capabilities within the same web application.

It would also be of interest to examine how different application architectures effect the effectiveness of object capability security. The implemented prototype utilized static type checking and a monolithic architecture, allowing it to apply techniques that are not available in a dynamically typed language or in a microservice architecture. These design decisions would then require a different set of OCAP-based techniques.

References

[1] K. Smith, A. Jones, L. Johnson, L. Smith, "Examination of cybercrime and its effects on corporate stock value," *Journal of Information, Communication and Ethics in Society*, vol. 17, no. 1, pp. 42–60, 2019.

[2] K. Mlitz. (2021). *Size of the cybersecurity market worldwide, from 2021 to 2026*. [Online]. Available: https://www.statista.com/statistics/595182/worldwide-security-as-a-service-market-size/. [Accessed: Oct. 24, 2022].

[3] Inc. OWASP Foundation. (2021). *OWASP top ten 2021* [Online], Available: https://www.hhs.gov/sites/default/files/owasp-top-10.pdf. [Accessed: Oct. 24, 2022].

[4] J. H. Saltzer, "Protection and the control of information sharing in multics," *Communications of the ACM*, vol. 17, no. 7, pp. 388–402, 1974.

[5] H. M. Levy, *Capability-based computer systems*. USA: Butterworth-Heinemann, 1984.

[6] B. W. Lampson, "Protection," *SIGOPS Operating Systems Review*, vol. 8, no. 1, pp. 18–24, 1974.

[7] D. Ferraiolo and R. Kuhn, "Role-based access control," *15th NIST-NCSC National Computer Security Conference*, 1992, pp. 554–563, doi: https://doi.org/10.48550/arXiv.0903.2171.

[8] V. Hu, D. Kuhn, D. Ferraiolo, J. Voas, "Attribute-based access control," *Computer*, vol. 48, pp. 85–88, 2015, doi: 10.1109/MC.2015.33.

[9] J. B. Dennis, E. C. Van Horn, "Programming semantics for multiprogrammed computations," *Communications of the ACM*, vol. 9, no. 3, pp. 143–155, 1966.

[10] J. S. Shapiro, J. M. Smith, D. J. Farber, "EROS: A fast capability system," *Proceedings of the 17th ACM Symposium on Operating Systems Principles*, 1999, pp. 170–185.

[11] A. S. Tanenbaum, M. F. Kaashoek, R. V. Renesse, H. E. Bal, "The amoeba distributed operating system – a status report," *Computer Communications*, vol. 14, pp. 324–335, 1991.

[12] K. Elphinstone, G. Heiser, "From L3 to seL4 what have we learnt in 20 years of L4 microkernels?" *Proceedings of the 24th ACM Symposium on Operating Systems Principles*, 2013, pp. 133–150, doi: 10.1145/2517349.2522720.

[13] Z. Wilcox-O'Hearn, B. Warner, "Tahoe: The least-authority filesystem," Proceedings of the 4th ACM International Workshop on Storage Security and Survivability, 2008, pp. 21–26, doi: 10.1145/1456469.1456474.

[14] C. Morningstar. (2017, May 7). What are capabilities? [Online]. Available: http://habitatchronicles.com/2017/05/what-are-capabilities/. [Accessed: Oct. 24, 2022].

[15] A. Conill. (2019). The bearer capability URI scheme. [Online]. Available: https://git.sr.ht/~kaniini/draft-conill-bearcapsuri-scheme/tree/22c458a95992e56ac41f1fff745855b14a811046/item/draft-conill-bearcaps-urischeme.txt. [Accessed: Oct. 24, 2022].

[16] C. Lemmer-Webber. (2019). Bearcap URIs. [Online]. Available: https://github.com/cwebber/rwot9prague/blob/908c5522720f0e3debad2c1578c28a984660ba05/topics-and-advancereadings/bearcaps.md. [Accessed: Oct. 24, 2022].

[17] N. Madden. (2021). Towards a standard for bearer token URLs. [Online]. Available: https://neilmadden.blog/2021/03/20/towards-a-standard-for-bearer-token-urls/. [Accessed: Oct. 24, 2022].

[18] A. Birgisson, J. G. Politz, Ú. Erlingsson, A. Taly, M. Vrable M. Lentczner, "Macaroons: Cookies with contextual caveats for decentralized authorization in the cloud," in Network and Distributed System Security Symposium, San Diego, CA, 2014.

[19] G. working group. (2020, July 10). Grant negotiation and authorization protocol. [Online]. Available: https://datatracker.ietf.org/doc/charter-ietf-gnap/01/. [Accessed: Oct. 24, 2022].

[20] C. Lemmer-Webber, M. Sporny, M. S. Miller. (2020, Dec. 29). Authorization capabilities for

linked data v0.3, World Wide Web Consortium Community Group. [Online]. Available: https://w3c-ccg.github.io/zcap-ld/. [Accessed: Oct. 24, 2022].

[21] D. Longley, M. Sporny. (2021, June 3). Linked data proofs 1.0, World Wide Web Consortium Community Group. [Online]. Available: https://w3c-ccg.github.io/ld-proofs/. [Accessed: Oct. 24, 2022].

[22] Agoric. (2021). Endo [Online]. Available: https://github.com/endojs/endo. [Accesed: Oct. 24, 2022].

[23] Agoric. (2021). ECMAScript spec proposal for ShadowRealm API. [Online]. Available: https://github.com/tc39/proposalshadowrealm. [Accessed: Oct. 24, 2022].

[24] C. Lemmer-Webber. (2021). Spritely: Social worlds await. [Online]. Available: https://spritelyproject.org. [Accessed: Oct. 24, 2022].

[25] C. Lemmer-Webber. (2021, July 18). Spritely goblins v0.8 released! [Online]. Available: https://spritelyproject.org/news/goblins0.8.html. [Accessed: Oct. 24, 2022].

[26] K. Varda. (2021). Cap'n proto: introduction. [Online]. Available: https://capnproto.org/. [Accessed: Oct. 24, 2022].

[27] Google. (2021). Fuchsia. [Online]. Available: https://fuchsia.dev. [Accessed: Oct. 24, 2022].

[28] J. A. Rees, "A security kernel based on the lambda-calculus," Massachusetts Institute of Technology, vol. 1564, 1995.

[29] M. S. Miller. (2011, Oct. 7). Bringing object-orientation to security programming. [Online]. Available: https://www.youtube.com/watch?v=oBqeDYETXME. [Accessed: Oct. 24, 2022].

[30] M. S. Miller, K.-P. Yee, J. Shapiro, Capability myths demolished, 2003.

[31] M. S. Miller, Robust composition: Towards a unified approach to access control and concurrency control. Baltimore, Maryland: Johns Hopkins University, 2006.

[32] M. S. Miller, C. Morningstar, B. Frantz, "Capability-based financial instruments," Proceedings of Financial Cryptography 2000, Anguila, BWI, 2000, pp. 349–378.

[33] J. Noble, S. Drossopoulou, M. S. Miller, T. Murray, A. Potanin, "Abstract data types in object-capability systems," IWACO, 2016.

[34] S. Peyton Jones, "A history of haskell: Being lazy with class," The Third ACM SIGPLAN History of Programming Languages Conference (HOPL-III), 2007.

[35] M. Finifter, A. Mettler, N. Sastry, D. Wagner, "Verifiable functional purity in java," Proceedings of the 15th ACM Conference on Computer and Communications Security, 2008, pp. 161– 174.

[36] S. Wlaschin. (2015). *Using types as access tokens, F# for Fun; Profit.* [Online]. Available: https://fsharpforfunandprofit.com/posts/capability-based-security-3/. [Accessed: Oct. 24, 2022].

[37] E. Evans, *Domain-driven design: Tackling complexity in the heart of software*. Addison-Wesley, 2004.

[38] C. Simpson. (2018). *Object capability discipline, Monte Project.* [Online]. Available: https://monte.readthedocs.io/en/latest/intro.html#object-capability-discipline. [Accessed: Oct. 24, 2022].

[39] P. Wadler and S. Blott, "How to make ad-hoc polymorphism less *ad hoc*," *Proceedings of the 16th Annual ACM SIGPLAN–SIGACT Symposium on Principles of Programming Languages, ACM*, 1989, pp. 60–76.

[40] T. Close, "Web-key: Mashing with permission," *Proceedings of Web 2.0 Security and Privacy,* 2008.

[41] N. Lœuillet. (2016). *Wallabag.* [Online]. Available: https://wallabag.it/en. [Accessed: Oct. 24, 2022].

[42] J. Schoning. (2019). *Espial.* [Online]. Available: https://github.com/jonschoning/espial. [Accessed: Oct. 24, 2022].

[43] S. SAS. (2005). *Symfony* [Online]. Available: https://symfony.com/. [Accessed: Oct. 24, 2022].

[44] T. Y. Team. (2012). *Yesod web framework.* [Online]. Available: https://www.yesodweb.com/. [Accessed: Oct. 24, 2022].

[45] T. P. Team (2017). *PureScript.* [Online]. Available: https://www.purescript.org/. [Accessed: Oct. 24, 2022].

[46] T. Bazaz, A. Khalique, "A review on single sign on enabling technologies and protocols," *International Journal of Computer Applications*, vol. 151, pp. 18–25, 2016.

[47] C. Michael, M. Gegick, S. Barnum. (2005, Sep. 12). *Complete mediation, The Cybersecurity; Infrastructure Security Agency*. [Online]. Available: https://us-cert.cisa.gov/bsi/articles/knowledge/principles/completemediation. [Accessed: Oct. 24, 2022].

[48] N. V. Limited. (2016). *Element.* [Online]. Available: https://www.element.io/ [Accessed: Oct. 24, 2022].

Commentary: The Czech Approach to Supply Chain Security in ICT

Veronika Netolická | **Masaryk University, Faculty of Social Studies, Department of Political Science and National Cyber and Information Security Agency, Cyber Security Policies Department, National Strategy and Policy Unit, Czech Republic**, ORCID: 0000-0001-8991-384X

Abstract

Supply chain security is one of the challenges many countries are currently addressing. As this topic is a national security prerogative, the systems for screening also vary. The Czech Republic is preparing a legislative framework to protect strategically important infrastructure from high-risk suppliers. This commentary focuses on the Czech Republic's progress in this area, particularly in the European context.

Keywords

national regulation, security, supply chain, the Czech Republic

Corresponding author:
Veronika Netolická, Masaryk University, Faculty of Social Studies, Department of Political Science and National Cyber and Information Security Agency, Cyber Security Policies Department, National Strategy and Policy Unit, Czech Republic;
ORCID: 0000-0001-8991-384X
v.netolicka@gmail.com

Cite this article as: Example: V. Netolická, "Commentary: The Czech Approach to Supply Chain Security in ICT," ACIG, vol. 1, pp. 210–216, 2022, DOI: 10.5604/01.3001.0016.0867

1. Introduction

The Czech Republic (CZ) is one of the last European countries to prevent and reduce dependencies on high-risk suppliers in its strategically important infrastructure through legal regulation. Despite the proactive Czech approach to this issue and its long-standing importance, CZ has not yet adopted a legislative framework to protect strategically important infrastructure from high-risk suppliers, making it among the last European Union (EU) member states to do (alongside Croatia and Ireland) [1]; as a result, it is one of the last countries not to use its prerogative to interfere in national security matters reserved for EU Member States. For this purpose, the Czech National Security Council has given authorisation to the National Authority for Cyber and Information Security (NUKIB) [2], which is the central administrative body for cyber security (including the protection of classified information in the information and communication systems and cryptographic protection). Before this mandate, CZ, like most European and EU countries, looked at supply chain security chiefly through the lens of the supply chain to 5th-generation (5G) networks. 5G networks, often referred to as technologies with the potential to become the backbone of the economy, were socially accepted as infrastructure that should be protected from misuse [3]. This narrow view which focused only on electronic communications networks—or only on one generation—has primarily (but not exclusively) shifted due to the changing security environment in Europe related to the war in Ukraine. The war in Ukraine has made it clear to the world, and Europe in particular, that supply chain security is important regardless of the sectoral divide or interlocking components for which we require this security. These interlocking components stand for confidentiality, integrity, and availability. Each has its irreplaceable function and needs to be equally protected to achieve the security of the whole security ecosystem. The supply chain of the information and communication infrastructure has the potential to disrupt each of these components. Screening of the supply chain, especially when it comes to infrastructure essential to the functioning of a state, must be the prerogative of each state underlying its technological sovereignty. Regarding the importance of this topic, this commentary focuses specifically on the supply chain of the country's information infrastructure at a strategic level, for which regulation is currently being drafted in CZ and on the basis of which CZ can exclude high-risk suppliers. The path of the CZ in this respect will be described in terms of developments in the international environment, with a primary focus on EU countries. On this basis, the current status quo and the basic ideas of the forthcoming mechanism are set out. Finally, the question of the role of the private sector in this topic is discussed.

2. The Czech Republic's path to the forthcoming regulation

CZ is one of the last remaining EU countries to approach supply chain regulation. However, few comprehensive approaches exist, even in Europe, with partial sectoral regulations prevailing. The German approach is the most comprehensive regardless of the sectoral scope. Its mechanism

gives the Federal Ministry of the Interior ex ante powers to restrict the use of a critical component if its operation would affect national security [4]. Germany is followed by countries such as Poland, Slovakia, Romania, Latvia, and Cyprus, which also have legislative powers to deal with supply chain security. Furthermore, countries such as Finland, Denmark, the Netherlands, France, Austria, Estonia, Belgium, Sweden, and Italy, have already adopted specific legislative measures to reduce the risks associated with high-risk suppliers in the electronic communications sector, specifically in response to the security of the 5G networks deployment [5]. The absence of any similar process places CZ among the countries with a lower resilience to these threats; however, CZ is otherwise one of the countries with the most advanced cyber security systems [6]. It should be noted that this is not a topic that is solely dealt with at the level of EU Member States. For example, countries such as the United States of America [7], the United Kingdom [8], or Japan [9] also have mechanisms to address supply chain security.

The creation of these screening mechanisms is the prerogative of each state, even if they are members of the EU. The topic of supply chain security impinges on national security issues that the member states themselves are best equipped to evaluate and ensure. However, the EU is not giving up its efforts to address the topic in a broad plenary of member states. The most significant input has been the 5G EU Toolbox publication in 2020 (Toolbox), which presents a set of measures to mitigate the risks associated with deploying 5G networks. The Toolbox's measures are non-binding, and the decision on the scope and implementation of each measure is thus left to the EU Member States themselves [10]. The next major step at the EU level is in the form of the Council Conclusions on ICT supply chain security [11]. The fact that this is happening under the auspices of the Czech Presidency underlines the importance and priority of this topic for CZ.

Although it may seem that CZ is lagging behind the other EU Member States, this is true only concerning the absence of legal regulation to date, not the lack of consideration of the importance of the topic as such or steps at the level of soft law. First and foremost, in 2018, NUKIB warned against using software and hardware from Huawei Technologies Co., Ltd. and ZTE Corporation [12]. There are also instruments currently in place that feed into the overall framework and will complement the forthcoming legislation. This includes, for example, the legal authorization of the state to screen i) foreign direct investments [13] and ii) the applicants of cloud computing providers for registration in the state catalogue [14]. The upcoming legislation should thus analogously supplement the state's authority in screening supply chains in its strategically important infrastructure, if CZ does not want to give up on ensuring its cyber security and national security.

Just as at the international level, the topic has been highlighted, and measures have been issued, so the first steps have been taken at the Czech national level. The 2018 NUKIB warning declared CZ's position to high-risk vendors, supported by the issuance of a series of recommendations known as the Prague Proposals: The Chairman Statement on cyber security of communication networks in a globally digitalized world

in 2019 [15] and developed by the Prague Proposals on Cyber Security of Emerging disruptive technologies (EDTs) [16], and the Prague Proposals on Telecommunications Supplier Diversity in 2021 [17].

In February 2021, as part of the original mandate, NUKIB prepared a white paper for regulating the supply chain verification mechanism solely for the electronic communications sector [18]. In this context, in February 2022, a non-legally binding Recommendation for assessing the trustworthiness of technology suppliers of 5G networks in CZ was issued [19]. These steps towards securing the supply chain to strategically important infrastructure in one individual sector have been followed by the latest decision of the National Security Council authorizing NUKIB to prepare a law to screen high-risk suppliers to strategically important infrastructure regardless of sectoral focus [20]. However, the fact that risks originating from the supply chain also affect sectors other than electronic communications is evident from the warnings issued by the NUKIB in 2022. The first of these warnings was against the use of smart meters from countries with untrustworthy legal environments [21]. The second warning was issued in the context of economic sanctions associated with the Russian Federation [22].

3. Principles and aspects of the forthcoming regulation

Although the legislation is still being drafted, NUKIB is already publicly commenting on the basic principles on which it wants to base the law, and on who will be affected by the regulation. An overall principle of NUKIB's work is openness concerning pending regulation and a high degree of involvement from the private sector and academia. In this context, expert public consultation has also been promised prior to the formal commentary process [23].

From the private sector's perspective, in terms of private entities or associations covered by the national regulation (the Act on Cyber Security), one of the most important questions was who would be affected by the regulation. NUKIB has been transparent in this respect, stating that the regulation will only affect the strategically important infrastructure of CZ. This term is then interpreted by the Act on Cyber Security as critical information infrastructure and operators of essential services—the set of entities is presented concerning the current legislation in force, but this set may change in the context of implementing the NIS2 Directive. The forthcoming legislation will respond to these changes to encompass only a defined set of entities, and not the extended range of entities that the implementation of the Directive will ultimately bring. Thus, the forthcoming law does not expand the group of persons and bodies under its remit but is based on a cross-section that is already familiar to the national format. Another important aspect is the substantive scope of the forthcoming regulation, i.e., what will be subject to scrutiny by the state. In this respect, too, it will not be the entire set of hardware and software subject to the administration of regulated entities. However, it will be those supplies that have relevance to national security. Thus, it only concerns predefined parts of the infrastructure [23].

The last accentuated principle according to which NUKIB wants to approach the upcoming regulation is to maintain the current approach to ensuring cybersecurity in CZ, according to which the administrator knows its system or network best and should therefore be responsible for its evaluation.

Thus, the mechanism will only affect critical parts of strategically important infrastructure while respecting the abovementioned principles. The preparation process is already underway with the involvement of the private sector and the academic sector.

4. Discussion: the role of the private sector

A change in the private sector's approach to increasing security need not always be driven by regulation or otherwise enforced by the state. Companies such as Palo Alto Networks [24] and Microsoft [25] have declared their interest in mitigating untrustworthy vendors. However, these remain exceptions, and even from the Czech perspective the private sector does not have a wholly unified approach. Thus, their position can be divided into two groups as a simplification. The first group follows the principle of short-term economic advantage at the cost of security risk in the medium and long term, while the second group advocates for efficient and sustainable security. As a result, the only adequate response is to shift this responsibility for assessing high-risk suppliers onto the state. Since states are best qualified at a strategic level to assess and evaluate these risks and they can also set (and oversee) a legal environment in which all equally require this element of security improvement, this demand from the private sector is also reasonable. Moreover, although the state is best equipped to make this assessment, despite the published and attributed cases that point to high-risk suppliers, there is not a sufficiently rapid change behind the mindset of using these suppliers, especially concerning the economic benefits they confer. In this regard, the state cannot afford to wait for a change in the attitude and mindset of the operators and administrators of its critical information infrastructure. Supply chain security is a complex issue, and CZ has shown that it endeavours to find a comprehensive solution. The actual result should become apparent in 2023.

References

[1] European Court of Auditors. (2022). *Special Report 03/2022: 5G roll-out in the EU: Delays in deployment of networks with security issues remaining unresolved*. [Online]. Available: https://www.eca.europa.eu/en/Pages/DocItem.aspx?did=60614 [Accessed: Oct. 17, 2022].

[2] Government of the Czech Republic. (2021). *Resolution: The National Security Council*. [Online]. Available: https://www.vlada.cz/assets/ppov/brs/cinnost/zaznamy-z-jednani/usn-41-22.pdf. [Accessed: Oct. 17, 2022].

[3] MPO. (2022). *Implementation and development of 5G networks in the Czech Republic towards the digital economy.* [Online]. Available: https://www.mpo.cz/assets/cz/e-komunikace-a-posta/elektronicke-komunikace/koncepce-a-strategie/narodni-plan-rozvoje-siti-nga/2020/1/Implementace-a-rozvoj-siti-5G-v-CR-EN.pdf. [Accessed: Oct. 17, 2022].

[4] CRS. (2022). *Supply Chain Act: Act on Corporate Due Diligence Obligations in Supply Chains.* [Online]. Available: https://www.csr-in-deutschland.de/EN/Business-Human-Rights/Supply-Chain-Act/supply-chain-act.html. [Accessed: Oct. 17, 2022].

[5] European Court of Auditors. (2022). *Special Report 03/2022: 5G roll-out in the EU: Delays in deployment of networks with security issues remaining unresolved.* [Online]. Available: https://www.eca.europa.eu/en/Pages/DocItem.aspx?did=60614. [Accessed: Oct. 17, 2022].

[6] NCSI. (2022). *Czech Republic.* [Online]. Available: https://ncsi.ega.ee/country/cz. [Accessed: Nov. 6, 2022].

[7] Federal Communications Commission. (2022). *List of equipment and services covered by section 2 of the Secure Network Act.* [Online]. Available: https://www.fcc.gov/supplychain/coveredlist. [Accessed: Oct. 17, 2022].

[8] UK Parliament. (2021). *Telecommunications (Security) Act 2021*. [Online]. Available: https://bills.parliament.uk/bills/2806. [Accessed: Oct. 17, 2022].

[9] A. Hiroshi. (2022). *Japan sets guidelines for protecting critical supply chains.* [Online]. Available: https://asia.nikkei.com/Spotlight/Supply-Chain/Japan-sets-guidelines-for-protecting-critical-supply-chains. [Accessed: Oct. 17, 2022].

[10] NIS Cooperation Group. (2020). *Cybersecurity of 5G networks EU Toolbox of risk mitigating measures.* [Online]. Available: https://ccdcoe.org/uploads/2020/01/EU-200129-Cybersecurity-of-5G-networks-EU-Toolbox-of-risk-mitigating-measures.pdf. [Accessed: Oct. 17, 2022].

[11] Council of the EU. (2022). *The Council agrees to strengthen the security of ICT supply chains.* [Online]. Available: https://www.consilium.europa.eu/en/press/press-releases/2022/10/17/the-council-agrees-to-strengthen-the-security-of-ict-supply-chains. [Accessed: Oct. 17, 2022].

[12] NUKIB. (2018). *Warning. National Cyber and Information Security Agency.* [Online]. Available: https://www.nukib.cz/download/uredni_deska/Varovani_NUKIB_2018-122-17.pdf. [Accessed: Oct. 17, 2022].

[13] MPO. (2022). *Implementation and development of 5G networks in the Czech Republic towards the digital economy.* [Online]. Available: https://www.mpo.cz/en/foreign-trade/investment-screening. [Accessed: Oct. 17, 2022].

[14] MV ČR. (2022). *EGovernment Cloud.* [Online]. Available: https://www.mvcr.cz/clanek/katalog-cloud-computingu.aspx. [Accessed: Oct. 17, 2022].

[15] Government of the Czech Republic. (2019). *Prague 5G security conference announced series of recommendations: The Prague Proposals.* [Online]. Available: https://www.vlada.cz/en/media-centrum/aktualne/prague-5g-security-conference-announced-series-of-recommendations-the-prague-proposals-173422 [Accessed: Oct. 17, 2022].

[16] NUKIB. (2021). *Prague proposals on cyber security of EDTs.* [Online]. Available: https://www.nukib.cz/download/Prague_Proposals_on_Cyber_Security_of_EDTs.pdf. [Accessed: Oct. 17, 2022].

[17] NUKIB. (2021). *Prague proposals on telecommunications supplier diversity.* [Online]. Available: https://www.nukib.cz/download/Prague_Proposals_on_Telecommunications_Supplier_Diversity.pdf. [Accessed: Oct. 17, 2022].

[18] Government of the Czech Republic. (2022). *Resolution: The National Security Council.* [Online]. Available: https://www.vlada.cz/assets/ppov/brs/cinnost/zaznamy-z-jednani/usn-33_2.pdf. [Accessed: Oct. 17, 2022].

[19] NUKIB. (2022). *Recommendation for assessing the trustworthiness of technology suppliers of 5G networks in CZ.* [Online]. Available: https://www.vlada.cz/assets/ppov/brs/cinnost/zaznamy-z-jednani/usn-33_2.pdf. [Accessed: Oct. 17, 2022].

[20] Government of the Czech Republic. (2021). *Resolution: The National Security Council.* [Online]. Available: https://www.vlada.cz/assets/ppov/brs/cinnost/zaznamy-z-jednani/usn-41-22.pdf. [Accessed: Oct. 17, 2022].

[21] NUKIB. (2022). *Warning against using smart meters from countries with untrustworthy legal*

environments. [Online]. Available: https://www.nukib.cz/download/uredni_deska/2022-05-30_varovani-smartmetering_final_1.0_podepsno.pdf [Accessed: Oct. 17, 2022].

[22] NUKIB. (2022). *Warning in the context of economic sanctions associated with the Russian Federation.* [Online]. Available: https://www.nukib.cz/download/uredni_deska/2022-03-21_varovani_rusti-dodavatele.pdf [Accessed: Oct. 17, 2022].

[23] NUKIB. (2022). *Increasing the supply chain security of the state's strategic infrastructure is in the interest of the Czech Republic.* [Online]. Available: https://www.nukib.cz/en/infoservis-en/news/1886-increasing-the-supply-chain-security-of-the-state-s-strategic-infrastructure-is-in-the-interest-of-the-czech-republic. [Accessed: Oct. 17, 2022].

[24] M. Coleman. (2020). *NIST highlights palo alto networks supply chain best practices.* [Online]. Available: https://www.paloaltonetworks.com/blog/2020/06/policy-supply-chain-best-practices. [Accessed: Oct. 17, 2022].

[25] Microsoft. (2022). *Supply chain security.* [Online]. Available: https://www.microsoft.com/en-us/research/project/supply-chain-security/publications. [Accessed: Oct. 17, 2022].

Russian Aggression against Ukraine as the Accelerator in the Systemic Struggle against Disinformation in Czechia

Ladislav Cabada | **Department of Political Science and Anglophone Studies, Metropolitan University Prague, Czech Republic**, ORCID: 0000-0001-9604-0987

Abstract

In the last decade Czechia's foreign and security policies were destabilised by the activities of external actors, with Russia in the leading role, and also by internal actors who followed the Russian and pro-Kremlin propaganda and disinformation campaigns and/or actively participated in such subversive activities. After 2015, within the set of crises and their securitisation, a disinformation network was developed in Czechia using social media and so-called 'alternative online media' for the dissemination of disinformation, misinformation, fake news and chain mails to spread these campaigns. As leading persons in the executive belonged to the disinformers, the government was not able to develop working strategies against the disinformation campaigns as the new hybrid threat until 2021. At the end of 2021, the new Czech government of Prime Minister Petr Fiala launched a new strategy regarding hybrid threats which contained disinformation. The one-year plan to establish a systemic platform for the struggle against such threats was challenged by Russian aggression against Ukraine. In this article, we analyse the development of the security eco-system in Czechia against these hybrid threats, specifically the acceleration and intensification of this activity after 24 February 2022.

Keywords

Czechia, disinformation campaigns, hybrid threats, propaganda, Russia

Corresponding author:
Ladislav Cabada, Department of Political Science and Anglophone Studies, Metropolitan University Prague, Dubečská 900/10, 10000 Prague, Czech Republic;
ORCID: 0000-0001-9604-0987;
ladislav.cabada@mup.cz

Cite this article as: L. Cabada, "Russian Aggression against Ukraine as the Accelerator in the Systemic Struggle against Disinformation in Czechia," ACIG, vol. 1, no. 1, pp. 217–234, 2022, DOI: 10.5604/01.3001.0016.0916

1. Introduction

Czechia belongs to NATO and the EU, it is one of the so-called 'new democracies'. Is also belongs to the Visegrád Group (V4) which is the cooperation of four Central European nations that share a similar historical experience and legacy. The most important negative legacy upon these countries is their former membership of the Eastern Bloc controlled by the Soviet Union. As such they were subjected to the totalitarian rule of communist ideology and communist parties. As an important modality of this Soviet control over Eastern and Central Europe the continuation of older Russian imperial projects is often discussed within the V4 political and societal discourse, such discussions are repeatedly confirmed by direct Soviet interventions against the liberalisation processes in Poland, Hungary and Czechoslovakia [1]. The effort to dismantle the (post-)communist framework and (re-)create the democratic order, liberal economy and civic society in Central Europe also covers the permanent solution of the geographic proximity of the region to Russia and security dilemmas related to Russian activities and geopolitical conceptions.

Naturally, not all actors in V4 societies accept and follow the above-mentioned axioms and stances, and within the polarised societies and political elite we can observe an *ad hoc* or even a stable 'Russian factor' influencing public debate and in some cases also public policies. As one of the recent analyses summarised:

> The CEE [Central and Eastern European] region represents a very unique space within the Euro-Atlantic area. Particularly because of its countries' several historical, linguistic, or ethnic ties to Russia, the narratives that are being circulated there often differ considerably from those observed in Western Europe or North America. As a result, the CEE region can be perceived as intrinsically more vulnerable to disinformation campaigns, especially because of the wider range of narratives that Russia can exploit there for such a purpose, including the Russian World, Slavic Unity or Ostalgia narratives. Simultaneously, the CEE region faces numerous deleterious trends that are favourable to Russian information warfare tactics. Most evident has been a continuous decline in citizens' trust in traditional media platforms, which are the least likely to be polluted with disinformation and misinformation [2, p. 18].

Regarding Czechia, the Russian 'factor' has been discussed since 1989 specifically with the continuation of the occurrence of the only partly-reformed and in many regards neo-Stalinist Communist Party of Bohemia and Moravia (*Komunistická strana Čech a Moravy*, KSČM), which was present in all parliaments elected between 1990 and 2021, and which participated actively in the existence of and policies promoted by at least two governments in this period – the left-centrist government led by Social Democrat Jiří Paroubek (2005–2006) [3, 4] and the government led by the oligarchic leader of the populist movement ANO 2011 Andrej Babiš (2017–2021) [5, 6]. Furthermore, the pro-Russian activities and stances of Presidents Václav Klaus and Miloš Zeman have often been discussed, in the case of Miloš Zeman his positive views on the authoritarian politics of

Chinese and Russian leaders became one of the symbols of the democratic backsliding of Czech (foreign) policy [7].

Together with Viktor Orbán, Miloš Zeman and Andrej Babiš became the most visible Central European political actors repeatedly undermining the joint EU and the general Western position towards Putin´s Russia. Furthermore, Prague became the hub of Russian intelligence and subversive activities in CEE, operated by the oversized Russian embassy. Their strategy also included the extensive use of disinformation (including cooperation with alternative media and other sources of disinformation campaigns), this was specifically so in the case of Zeman. Zeman belonged to the small group of European politicians who developed extensive and friendly relations with Vladimir Putin after the Russian annexation of Crimea, he continually cast doubts upon Western sanctions against Russia [8], and he strongly promoted the participation of Russian state-owned firms in the tenders for nuclear reactors in the new Czech power plants. All this brought him into conflict not only with the right-wing political parties in the country, but also with Czech intelligence, above-all BIS (*'Bezpečnostní informační služba'* – Czech Security Information Service) and its Director Michal Koudelka. Repeatedly, Zeman questioned the activities of BIS and the qualification of its director, for the public he labelled BIS 'dabblers' (in Czech 'čučkař'). Despite the government´s repeated proposal to promote Koudelka to the rank of general, Zeman consistently rejected this proposal, including in autumn 2022 [9].

Only after April 2021 and the so-called 'Vrbětice case'[1] did the situation (partly) change. Indeed, President Zeman continued in his pro-Russian activities till 24 February 2022, the beginning of Russian aggression against Ukraine. Furthermore, his declarations about his misunderstanding regarding the intentions of Putin, although limited in time and scope, did not give the impression of a real turnabout. Specifically, we have to stress the role of Zeman´s main advisor, Martin Nejedlý, and his direct ties to the leading persons of Putin´s regime [11].

As showed, in Czechia there are strong and active political actors with a pro-Russian inclination using a colourful set of instruments and measures for the dissemination of fake news and spreading of (pro-) Russian propaganda disinformation. In the case of Czech president Miloš Zeman, these activities began immediately after his entrance into office, in the case of Andrej Babiš, they began as early as the so-called 'migration crisis' in 2015. Both politicians, along with the leader of the radical, and in some issues extreme-right, political party Freedom and Direct Democracy (*Svoboda a přímá demokracie*, SPD), Tomio Okamura, addressed the part of the Czech population prone to disinformation campaigns using the ´politics of fear´ strategy [12]. Specifically, between 2017 and 2021, these actors cooperated in strengthening such a disinformation discourse – Babiš became the prime minister and his minority government often depended on the support of the president as well as SPD deputies in the parliaments. Naturally, in such an environment the struggle against disinformation, and more generally for (cyber) security, became mere rhetorical adornment without any clear content.

Furthermore, the set of new crises – COVID-19 and the pandemic, the economic crisis and energy insecurity – strengthened the polarisation

1 —— The Vrbětice case presents the most intensive and visible Russian interference in Czechia. In October and December 2014 there were two series of detonations in the munition storehouse in Vrbětice, killing two people. In April 2021, the Czech government declared that based on the BIS investigation, the main directorate of the general staff of the Armed Forces of the Russian Federation (GRU) was responsible for these acts of sabotage. While President Zeman continued his pro-Russian activities, the government carried out a set of measures including the expulsion of dozens of Russian diplomats (and/or agents) from Czechia. Nevertheless, several analyses reflect the delayed response of Czech institutions including the delayed announcement of the BIS investigation. Furthermore, evidence exists that the chair of the Social Democratic Party and at that time minister of internal affairs as well as minister of foreign affairs, Jan Hamáček, was thinking about the import of Russian anti-Covid vaccines Sputnik-V in exchange for secreting the role of Russia in the terrorist attacks in Vrbětice [10].

of Czech society as well as the activities of negative (so-called 'desolates') and extremist actors in Czech society and politics. As the recent annual report of the Czech Security Information Service notes:

> In the course of the COVID-19 pandemic, disinformation narratives gained increased popularity in Czech society, as disinformation spread with growing speed on social media. In 2021 the most prevalent narratives of the online disinformation space focused on COVID-19, vaccination and pandemic-related restrictions [...] One of the main sources of the information shared among supporters of the COVID-denial movement were articles published by disinformation media. BIS noted some disinformation narratives originating from foreign-language websites too. In the course of 2019[2] some COVID-denial activists underwent a slow radicalisation in terms of opinion and rhetoric, however, their beliefs and protest activities failed to appeal to most of society. Having a mostly symbiotic relationship with the COVID-denial movement, pro-Russian activists used COVID-19 as a vehicle for spreading conspiracy theories, disinformation and pro-Kremlin propaganda [13, p. 16].

2 —— We assume that there is a mistake in the report, and the years 2020 and 2021 were meant.

The aim of this short analytical article is to examine the most important changes in the Czech discourse on disinformation after 24 February 2022, i.e. after the beginning of Russia's war against Ukraine. Furthermore, our goal is also to discuss the threat of disinformation campaigns to the stability of the democratic order and security of Czechia, and above all the recent political and public discussion about proper and effective measures against disinformation campaigns.

The Czech government of Petr Fiala, the main actor in the changes (or attempted changes), existed for only two months, at the time when Russia's aggression started. Having the newly presented government manifesto, the Czech government was challenged to change, modify or supplement its plans. As long as we focus on an ongoing and unfinished political process, we cannot fully use scholarly publications, but rather a mixture of sources including reports, policy and position papers and articles by journalists. Methodologically, our article is rooted in the process tracing method and discourse analysis. We will analyse firstly the security eco-system in Czechia, especially the challenges of new hybrid threats. Specifically, we will focus on cyber security and disinformation which comprise the distinctive segments of new threats. Secondly, we will focus on legal, political and institutional changes in the Czech security eco-system reflecting the individual domain of struggle against disinformation.

2. The Czech security eco-system and the struggle against the hybrid treats

Czechia is part of the Western security community, where membership of NATO and the EU is the most important delimitation. Compared with two other V4 partners, Hungary and Poland, and similar to Slovakia, Czechia does not strongly prioritise the so-called 'national interest' in its strategic documents, but rather belonging to the international security community.

The Czech strategy documents generally work with the issue of hybrid threats, a specific impetus for the development of this partial issue was the Russian occupation of Crimea.

> The Czech Republic has a system of security documents (strategies and related plans, concepts etc.) which are basically hierarchically organised and interconnected. Although they show a departmental approach, they are generally designed to have an impact on the entire spectrum of public administration [...]. The Security Strategy of the Czech Republic is the document with the highest political (not legal) force in the Czech security sphere. Its latest version from 2015 contains important passages on hybrid warfare, which was a reaction to the then developing crisis in Crimea and eastern Ukraine. In 2016, a unique document "outside of the hierarchy" of the strategies and concepts outlined above was also prepared, namely the National Security Audit, which included a chapter on hybrid threats. This chapter was under the responsibility of the Ministry of Defence, while the entire document was under the responsibility of the Ministry of Interior and was approved by the government. In 2021 the government then approved a specialised strategy document prepared by the Ministry of Defence called "National Strategy for Countering Hybrid Operations" [14, pp. 344–345].

A very inclusive manner is used to focus on hybrid threats, reflecting their changeable nature, modalities and mutual interconnection. 'The Security Strategy of Czechia works with the concept of "hybrid warfare methods", which, according to the Strategy, combine "conventional and unconventional military means with nonmilitary tools (propaganda using traditional and new media, intelligence disinformation actions, cyberattacks, political and economic pressure, sending unmarked members of the armed forces"' [14, p. 345].

Jan Daniel and Jakub Eberle [15, p. 907] analysed the role of the main relevant actors, such as bureaucrats, scholars, journalists, think-tanks and non-governmental organisations, shaping the paradigmatic turn in the Czech security discourse. Summarising the result of the analysis, Mareš et al. note: 'The factor of the combination of conventional and unconventional armed forces in combat operations was thus neglected, and nonmilitary and nonviolent (or at least less violent) forms of conflict, such as propaganda, embargoes or attacks in cyberspace, which are intended to destabilise society, came to the fore' [14, p. 349].

3. The challenge of disinformation campaigns

As already mentioned in the introduction, one of the most challenging problems from the group of hybrid threats is the disinformation campaigns. Here the situation in Czechia partly follows the more general trends in Western societies affected by the wave of neo-illiberal populism [16], nativism, cultural backlash [17] and in some cases even democratic backsliding and de-democratisation [18, 19]. The liberal political and media mainstream, as well as the scholars in social sciences, reflect the disinformation as

a serious hybrid threat based on some partial events, but also based on the continuous development of the disinformation strategies. Among the first events that provoked a debate about the impact of disseminating disinformation were the beginnings of the Ukrainian crisis in 2013, the pro-Brexit campaign in the United Kingdom, the possible influencing of the US presidential elections in 2016 [20] and, from more recent issues, the disinformation campaigns concerning socially questionable issues such as Black Lives Matter, the European Green Deal, mass migration from the MENA region to Europe, and obligatory vaccination against COVID-19. In the case of Czechia, specific mention needs to be made of disinformation in both the direct Presidential elections in 2013 and 2018 by the team surrounding the candidate, and later president, Miloš Zeman. In 2018 analysts also demonstrated the use of disinformation against Zeman before the second round of elections. As the main sources of disinformation, the analysis detected the information web Aeronet.cz [21], hosted in the Netherlands by an Indian client to be providing services to Russian companies and spreading proKremlin narratives [22, p. 42].

The Czech authorities formally responded on the strengthening disinformation. Based on the security document mentioned above as well as the lessons learned from the presidential campaigns, the Centre Against Terrorism and Hybrid Threats (*Centrum proti terorismu a hybridním hrozbám*, CTHH) was established within the Ministry of Internal Affairs. As declared by the then minister of internal affairs, Social Democrat Milan Chovanec, the aim of the group with nine members at maximum, is the expert, analytical and communication focus on mainly hybrid security threats such as terrorism, attacks against hard and soft targets, and security aspects of migration, extremism, mass events, breach of the peace and various criminal offences, as well as security aspects of disinformation campaigns related to the internal security of the nation. The directors of CTHH repeatedly stressed that the role of the institution is not censorship or even punishment, but mainly communication with the public, including information about the most visible disinformation sources.

Since the very beginning of its existence, the main enemy of CTHH has been President Zeman. Prime Minister Babiš also dismissed CTHH and took actions against its activities. In July 2021 during a debate within the National Security Council (*Bezpečnostní rada státu*) about the report on disinformation prepared by CTHH, the prime minister stopped it, asking (for time) to rework the material. As insiders shared with the media, Babiš rejected the recommendation to centralise the struggle against disinformation under the State Department. Paradoxically, the last meeting of the National Security Council before the parliamentary elections in October 2021 was cancelled (so the reworked report was not discussed). The main reason was Babiš´s electoral campaign, when he visited Hungarian prime minister Orbán, stressing that the main information issue was migration. As the media has pointed out, specifically with regard to migration issues, Babiš is Czechia's most prominent disinformant [23].

Generally, under the government of Babiš the activities of the CTHH were strongly restricted (financially, personally and also regarding preference of other priorities). As the recent director of the Centre Benedikt Vangeli notes:

> Our role was limited on what was allowed to us. We were devoted primarily on monitoring and internal security analyses incl.[uding] disinformation. None of the other institutions on the state administration do that; and none of the institutions provide this analysis systematically, as yet [...] We were trying to offer our help in the period of the previous government [of Prime Minister Babiš – quoted by Ladislav Cabada], but interest was minimal.

As Vangeli concludes: 'With the arrival of the new Minister of Internal Affairs, Rakušan, we can be more open' [24].

In the situation of limited government activities against disinformation campaigns, their detection as well as the identification of prominent disinformation sources became the activity of several non-governmental organisations, such as the Prague Security Studies Institute (PSSI), the Czech Elves (*Čeští elfové*), Manipulátoři.cz, Demagog.cz, project 'NELEŽ'[3] as the Czech franchise of the Global Disinformation Index, and many others. As Radek Bartoníček stressed, the volunteers and NGOs had already been warning the public for several years before the lies were widely disseminated, but only during the government led by prime minister Andrej Babiš did the state institutions fight against such disinformation exceptionally hard. All this despite the fact that such lies often endangered health and lives, for example during the Covid-19 pandemics. As Bohumil Kartous from the NGO Czech Elves noted, 'the government could hardly act effectively when Babiš himself positively reflected positively upon the disinformation chain mails and disseminated this fake news in the electoral campaign' [23].

3 —— In Czech translation two different, but interconnected, meanings – 'do not be lying', but also 'not-lie'

Czech academic institutions also developed an institutional framework for the struggle against disinformation. With the support of the Technological Agency of Czechia, in 2020 the team at Charles University in Prague, led by prominent journalist and scholar Václav Moravec presented the portal Infomore.cz, devoted to the disinformation related to the COVID-19 pandemic. Based on this activity, in the beginning of February 2022 (only two weeks before the Russian attack on Ukraine) Moravec presented the follow-up project of the Central European Digital Media Observatory, interconnecting the stakeholders from Czechia, Slovakia and Poland. The Centre belongs to the group of eight similar centres included the European Digital Media Observatory network [25]. A similar focus on disinformation campaigns and general media literacy and civic education has led to the international project Media Literacy Observatory for Active Citizenship and Sustainable Democracy (MELIA), supported by the Danube Transnational Programme. The Czech Republic Metropolitan University Prague as well as the NGO Edhance Plus have implemented the project [26].

Based upon the above mentioned project, as well as many others, there exists a plethora of scientific reports and journalist's articles including the repeatedly updated lists of prominent disinformation sites. Not only are the primarily pro-Russian websites located on these lists, but also the actors who generally contribute to the creation of the alternative media network. Such actors include online media and individual web pages. The flagship within this group is *Parlamentní listy* – an information portal which spreads fake news and disinformation. What they all have in common is their ability to flexibly interconnect different issues, for example: migration, the pandemic, the energy crisis, and the Russian invasion of Ukraine.

Nevertheless, as the recent annual report of BIS stressed:

> The disinformation scene in 2021 had an interconnected but centralised structure, disinformation and conspiracies were spreading mainly within the disinformation scene itself. Disinformers directed their activities mainly towards persons experiencing difficult life situations or disconnected and frustrated individuals. A key element of their motivation was to make financial profit, e.g. in the form of financial donations from their audience [13, p. 16].

A similar attitude was presented by the advisor to the government for the strategic communication and fight against disinformation, Dominik Presl, who specifically stressed the interest of the disinformants in advertisement and the necessity to prevent any orders for advertisements made by state-related institutions at such sites. He also confirmed that education about the media was weaker in Czechia than in the EU-15: 'Compared with the West, mainly with Scandinavia, the problem of disinformation campaign[s] is much bigger in Czechi Republic, we do have [a] much stronger eco-system of disinformation websites and also the disinformation influencers, thus the impact is bigger [4].

4. Russia as the key actor of hybrid threats and disinformation campaigns

On 18 October 2022, the Czech Security Information Service (BIS) presented its annual report for the year 2021. The relatively short document (about 30 pages) reflects the most important challenges and threats Czechia is facing. Reading the document, structured into several chapters, the part 'Intelligence and Subversive Activities Targeting the Czech Republic' has to be specifically stressed as reflecting cybersecurity issues generally and particularly disinformation activities. The chapter is divided into five sub-chapters with titles that very clearly show the main actors and matters in the cybersecurity area; the titles are 'Russia', 'China', 'Iran', 'Cybersecurity', and 'Activities contrary to the principles of democracy'. While the three subchapters dedicated to the activities of foreign governments against Czechia make up the bulk of diverse subversive operations, the last two mentioned parts specifically focus on defined 'single-issue' activities considered important components of post-modern hybrid warfare. The sub-chapter Cybersecurity describes cyber-attacks as state-sponsored cyber-espionage including: 'the spreading of surveillance malware, exfiltrating data from compromised victims or controlling of some attacker´s infrastructure in other countries.' It also states that 'the attacks were committed by a range of cyber actors (involved in either espionage or crime, including ransomware attacks and crypto mining)' [13, p. 13–14].

As the report further notes:

> Disinformation platforms often displayed narratives consistent with [the] interest of foreign powers (namely the Russian Federation and the People´s Republic of China). However, numerous disinformers worked on

> their own initiative and their activity was only loosely inspired by these narratives. Some representatives of anti-establishment and populist political entities took part in spreading the narratives on disinformation platforms [13, p. 16].

Such a statement is in concordance with recent scholarly research. We will present here two such analyses, one aimed on the Czech case, the second one at the V4.

As regards the (pro-)Russian propaganda, interesting outcomes were presented, for example, in an article written by Miloš Gregor and Petra Mlejnková. The authors analysed four selected model cases – Parlamentní listy (parlamentnilisty.cz), AC24 (ac24. cz), *Svět kolem nás* (svetkolemnas.info) and Sputnik (cz.sputniknews.com). The analysis confirmed the majority of manipulative techniques, such as blaming (pinpointing the enemy responsible for the event or situation), demonisation (dehumanisation of the opponent), appeal to fear, fabrication (presenting false information as true), labelling and relativisation [27, p. 546–557]. As they noted in their concluding remarks, the disinformation campaigns are not (only) rooted in lies:

> The disinformation campaigns in the Czech Republic do not necessarily need to lie to the audience [...] It is about choosing a narrative from the selected topics and stories and combining it with carefully chosen manipulative techniques. This could be seen as regards the conflicts in Ukraine and Syria. Data shows that the importance of the topic and Russian activities in both countries were relativized in order to persuade readers that nothing serious was happening there. In the case of Syria, the relativization of Russian activity was done by putting it in obvious and false counterbalance to US activities somewhere else and at a different time [...] manipulation was more sophisticated than obviously serving Russia as the only partner and alternative to the West. In our findings, Russia was mentioned in about a third of the news articles, but mostly in a neutral manner. Articles portraying Russia positively and negatively were about equal in number. In general, disinformation campaigns are more about redirecting blame onto others and lowering the level of trust in governments, elites, and established media within the general public [27, p. 559].

Lilla Sarlota Bánkuty-Balogh extended the analysis of disinformation campaigns on the V4. She identified five metanarratives: '(1) growing Russophobia in the West; (2) the preparation of a war against Russia by the US and NATO; (3) the United States seeking global hegemony; (4) the establishment of a system of Post-/NeoAtlantism by dividing Europe; and (5) the envisioned collapse of the European Union' [28, p. 187–190]. The author specifically analysed the disinformation campaigns in favour of Russia summarising that:

> The main focus of supposed western disinformation campaigns against Russia involved the Skripal and Navalny cases, insinuating Russian involvement in the United States presidential elections and rewriting or

> falsifying Second World War history in a way that depicts Russia as an aggressor. Narratives identified from the articles concerning Russia were overarching for all V4 countries; however, differences could be found in the frequency of mentions among them on a countrytocountry basis [...] Czech and Slovak language articles frequently featured alleged FBI and CIA involvement in manipulating local media to spread antiRussian sentiment with particular focus on the Skripal case [28, 178–179].

Specifically regarding Ukraine she notes that:

> Common narratives for the four countries included the hypothesised role of the United States in organising the Euromaidan, a wave of demonstrations in Ukraine which began in Maidan Nezalezhnosti (Independence Square) in Kyiv, later on followed by the Crimean crisis. The supposed rationale of the US was the provocation of Russian involvement in the Crimean crisis and ultimately the incitement of economic sanctions against Russia as well as nurturing Russophobia in neighbouring countries. The MH17 disaster was also linked as a planned incident to punish the Russian Federation for the annexation of Crimea. In both the Hungarian and Czech language news, the Nagornocrisis has been connected to Ukraine as well, with the alleged support of Kyiv to Azerbaijan during the conflict [28, p. 179].

As stressed in the introduction, Russia develops extensive subversive activities in Czechia, and the key Czech institutions ignored this fact until 2021, or even cooperated in such subversive activities, including the disinformation campaigns. Nevertheless, the analysis presented above also showed that in many cases, the disinformants are not directly connected to Russian officials, neither is there found any evidence of direct payments from Russia for these actors. This is also valid for the organisers of anti-government and pro-Russian demonstrations in autumn 2022 which included colourful mixtures of domestic (from the radical-right SPD party, the Communist Party, the ultraconservative and national-populist scene, and pro-Russian pan-Slavonic streams among others) and also international actors (representatives of the German Alternative for Germany party or the Serbian ultra-nationalist scene). Repeatedly, the organisers and speakers stressed that the protest should be transformed into riots and finally towards the violent takeover of power [29]. Such a concentration of anti-system and extremist actors proves a continued development in the last decade. To quote from the annual report of BIS, stressing that in the year 2021:

> It has been confirmed that the disinformation scene in the Czech Republic makes pragmatic use of any new topics which arouse a strong emotional reaction [...] At the end of the year, the most prominent activists organised a series of protest meetings, which were also attended by individuals from the anti-establishment milieu (including extremist and militia groups). As a result, the COVID-denial movement facilitated the spreading of radical and extremist views in society [13, p. 15–16].

5. New government plans and the effect of the Russian War against Ukraine

As repeatedly stressed, before the end of 2021 important Czech political actors underestimated and downplayed the seriousness of disinformation campaigns undermining and weakening democratic order in the country. Furthermore, some of the leading politicians participated in the use of disinformation campaigns. The new government led by Prime Minister Petr Fiala which was comprised of five political parties with rather colourful ideological orientations[4] stressed in its Policy Statement presented on 7 January 2022 the intention to struggle against the hybrid threats and disinformation. According to the document's chapter entitled Security:

- By the end of 2022, we will establish the "National Security Adviser" at the Office of the Government of the Czech Republic as the supra-ministerial coordinator for hybrid threats, disinformation and other serious supra-ministerial security issues. Therefore, the Office of the Government will have a platform for coordination and communication between security policy entities to ensure closer cooperation between intelligence and security forces and effective action against disinformation and hybrid threats.

- We will introduce rules for more transparent functioning of media: listing of publishers, owners, major sponsors and publishing of fina-ncial statements.

- We will adapt to the development of the security environment and promote a more professional approach to defence in the information space. Defence against disinformation must be fast and scalable. Following examples from abroad, we will prepare legislative and non-legislative measures that will allow us to better defend against harmful disinformation without compromising the freedom of expression [30].

As the government's statement declared, the Office of the Government (*Úřad vlády*) should become the incubator in the struggle against disinformation. In parallel, the government promised to create the position of the coordinator of hybrid threats, disinformation and other serious security risks. In January 2022, the Czech minister of internal affairs Vít Rakušan assumed that the coordinator would be declared by the end of that year. Regarding this issue Rakušan declared that 'Czech Republic needs the coordination of the security community, should it be in the matter of disinformation, intelligence, or the strategic communication of the state' [24].

Indeed, Russian aggression against Ukraine accelerated the preparatory phase and one month after the beginning of the war the government agreed upon the person to fill the role of new government coordinator for the media and disinformation, Michal Klíma. Klíma had worked as a media manager and also acted within several NGOs engaged in media ethics and education, such as the Czech National Committee of the International Press Institute; he also acted as the Chairperson of the Foundation for Holocaust Victims (*Nadační fond obětem holocaustu*). The

4 —— The government comprises of two liberal conservative parties – the Civic Democratic Party (Občanská demokratická strana, ODS) and Tolerance, Responsibility, Prosperity 09 (TOP 09) and Christian-Democratic party (KDU); these three actors created the electoral alliance Together (Spolu) for the elections. The second alliance was created by the centrist Mayor and Independent Candidates movement (STAN) and the Czech Pirate Party.

new coordinator serves as the advisor to the government and is subordinate directly to the prime minister. As Klíma stressed in his first statements, the government will specifically examine the purpose of disinformation campaigns [24].

The coordinator for media and disinformation is included in the broader framework of the prime minister´s advisors for security issues. Former diplomat, Tomáš Pojar, became the leader of this group and is expected to become the new national security coordinator based on his Israeli (Pojar served as the Ambassador to Israel) or U.S. experience. Alongside Klíma another important role is held by the former elite soldier Petr Matouš, who – along with other activities – served in Afghanistan. He came to work at the Office of the Government during the previous government´s term as the coordinator of the newly formed group struggling against hybrid threats [31]. While Klíma has to focus mainly on media and disinformation, Bečvář has to study the hybrid threats more comprehensively and holistically. As he stressed: 'Disinformation presents only a small fragment from the mosaic we focus on'. As he further noted, the hybrid threats got a media shortcut in disinformation and partly also in cyber-attacks, but the scope is much broaderAs he further noted, the media underplay the significance of hybrid threats, particularly of disinformation, but also of cyber-attacks. The scope of the threat is in reality much broader than is brought to public attention. Nevertheless, as one of the important prerequisites for the successful implementation of the Action Plan for the National Strategy for the Hybrid Activities Confrontation he announced the legislative eco-system for the struggle against disinformation [24].

This short overview of the new governments' plans shows that the general intention was to build up a new institutional framework for the struggle with hybrid threats during 2022. Nonetheless, Russian aggression against Ukraine caused the acceleration of this process, and also brought about the need for some extraordinary measures against the disinformation campaigns.

On 25 February 2022 the Czech internet society CZ.NIC[5] decided to blockade eight disinformation websites[6]. Such an unprecedented decision was made after consultation with the security services, it was also based on recommendations from the Czech government. As the Executive Director of the society, Ondřej Filip declared such an extraordinary measure as a response to the military attack of the Russian Army against Ukrainian sovereignty as well as to the disinformation campaign that accompanied and still accompanies this attack. As Filip further fully stressed, the blocked websites spread disinformation related to the Russia-Ukraine conflict. Government representatives refused to comment on the issue or on the legal framework for the blockade. Nevertheless, at the press conference organised by the government their position was made clear: 'We are in a disinformation war, and I will not divulge the strategy we follow' said Prime Minister Fiala [32].

Less than one week later, on 1 March 2022, Czech mobile operators joined the measures against the dissemination of Russian propaganda in Czech cyberspace when they blocked six disinformation webpages in the Czech language[7]. As the Association of Mobile Network Providers (*Asociace provozovatelů mobilních sítí*) stressed, this action satisfied the

5 —— CZ.NIC operates the domain register with the Czech national ending '.cz'.

6 —— Aeronet.cz, Protiproud.cz, Ceskobezcenzury.cz, Voxpopuliblog.cz, Prvnizpravy.cz, Czechfreepress.cz, Exanpro.cz and Skrytapravda.cz.

7 —— cz.sputniknews.com, Cz24.news, Nwoo.org, Slovanskenebe.com, Svobodnenoviny.eu a Zvedavec.org.

appeal from government and the National Centre for Cybernetic Operations (*Národní centrum kybernetických operací*, NCKO)[8] which operated within one of the Czech security services, namely Military Intelligence (Vojenské zpravodajství, VZ) As the President of the Association Jiří Grund declared:

> The appeal of state institutions stressed the extraordinality and urgency of the situation embodied in the ongoing aggression of the military forces of the Russian Federation in the territory of Ukraine. The threat for the Czech Republic´s security presents the dissemination of untrue and misleading information serving to the aggressor to deceive and manipulate Czech citizens with the aim to justify and endorse the recent aggression against Ukraine.

The steps of the mobile network providers followed their previous activity, namely the blockade of Russian state television programmes.

Paradoxically, the described (extraordinary) blockade of disinformation, internet media and content happened only a few days after the media inquiry focusing on one of the new government's goals, namely the struggle against disinformation. One week before the Russian attack on Ukraine the internet media website Aktualne.cz concluded that they could observe the first signs of the forthcoming attack. In response the new government launched several activities against the fake news and disinformation. Among others, the Ministry of Internal Affairs was preparing the campaign trying to calm down polarised Czech society. On the other hand the inquiry reflected the recommendations to the government to dedicate more determination and more people to this issue [24].

The blockade discussed above does not have any clear support in the Czech legislature, which does not recognise the term (and crime) of disinformation or propaganda. The use of disinformation or propaganda can only be prohibited and punished when related to different types of criminal acts, namely: interference in an individual´s rights, slander, calumny, defamation of nation, race, ethnic or other group of persons, instigation of hate against a group of persons or limiting their rights and freedoms, scaremongering news dissemination, endorsing a criminal offense, instigation of a criminal offense or declaring the declaration of liking towards movements oriented to suppression of human rights and freedoms [33].

As Petr Prchal noted in his comprehensive legal analysis of this extraordinary measure: 'only in April 2022, the server iROZHLAS.cz succeeded to get and verify the information that NCKO asked the director of CZ.NIC for the blockade,'. Prchal labels such an approach of state institutions as miserable. He also stresses that one of the main challenges is to define the disinformation [34]. On the other hand, Dominik Presl, advisor to the government for the strategic communication and fight against disinformation, believes that a definition of disinformation exists, namely 'untruthful information which is wilfully disseminated' [35].

The extraordinary blockade of selected disinformation websites was terminated three months after it began, i.e. on 25 May 2022. An exceptional case was the domain Aeronet.cz, where the CZ.NIC declared the absence of correct information about the possessor in the domain register. As the spokesperson of CZ.NIC noted, the continuation of the blockade would

8 —— NCKO operates within one of the Czech security services, namely Military Intelligence (Vojenské zpravodajství, VZ)

be possible only based on a court or police order, or the order of another competent state institution [36].

As the above-mentioned extraordinary blockade showed, (not only) the Czech legal system is challenged with the fuzzy definition of disinformation and above all the necessity to prevent any censorship. This 'optimism' became much more sober after several declarations given during March and April 2022 by some government members, including the minister of justice Pavel Blažek and the government coordinator for media and disinformation Michal Klíma, who mentioned the work on the new act against disinformation. On 13 April 2022 the Director of Legislative Division at the Office of the Government, Jan Večeřa, denied that the new Act would be in process of preparations. The non-existence of such activities was also declared by the spokesperson of the Czech Pirate Party, Veronika Šmídová. The Party is represented in government by the party chairman and minister for digitalisation, Ivan Bartoš, as well as the minister for legislating, Michal Šalamoun. Also, the minister for internal affairs Vít Rakušan stressed that 'the repressive steps in the form of new legislation present the utmost solution that is recently discussed only theoretically' [37].

Such a rational approach confirmed that the specialised legislature against disinformation is an extremely challenging issue. As Prchal notes: 'It is obvious that the problem society and [also] the law as the regulatory instrument are confronted with, is the formation of so-called social bubbles that might – in the utmost case – present a security risk' [34].

6. Conclusions

On 26 October 2022 the District Court in the Central Bohemian city Kladno sentenced two visible members of the Czech disinformation scene – Tomáš Čermák and Patrik Tušl. As the public prosecutor mentioned, both defendants presented the video on social networks in August 2022, where they disparaged Ukrainians as refugees. As the prosecutor stressed, both defendants used the disinformation as well as threats in the past, also against the representatives of the Czech Chamber of Medicine Doctors during the COVID-19 pandemic. Based on such continuous activities they were in custody between August and October 2022. Based on the court's decision, Čermák was given a six-month custodial sentence and Tušl ten-months, because they committed defamation of nation and instigation of hate [12]. This lawsuit is a precedent d showing that Czech justice is able to use recent legislature to punish disinformation campaigns. On the other hand, it also shows that in all cases the police, and later the court, need to find clear intent. Furthermore, it is also visible that only repeated inappropriate activities – and possibly the extraordinary situation caused by Russian aggression against Ukraine – brought state institutions towards a solution which was both relatively resolute and fast.

Thus, we have to stress that recently we cannot observe any change in the struggle against disinformation campaigns in Czechia. After the three-month long blockade the situation returned to its previous state. Here we fully agree with Martin Fendrych, who noted in October 2022:

> Beginning with the invasion of Ukraine, Russian state media and the dissemination of their content were prohibited in EU-member states. Indeed, the impact on the Czech case was not significant. Paradoxically, in the Czech Republic Russian propaganda is mainly spread by Czechs and Czech disinformation websites, Facebook groups, and Czech disinformation influencers. Without this Russian disinformation "fifth column" Russian propaganda would have a much smaller effect [39].

Such a sceptical assessment was also confirmed in the most recent annual report of BIS, despite the fact that this report officially reflects the situation in 2021; being published only in October 2022, it also surely included the impulses given during 2022. As the report points out:

The most prominent element in the disinformation ecosystem was websites which either contained disinformation or manipulated true information. Due to their far-reaching popularity, these websites had an impact on the rest of the alternative media scene and their articles were widely shared on social media [...] The disinformation eco-system is interconnected. Disinformation continues to spread primarily on Facebook, disinformation websites and through chain e-mails. A part of the anti-establishment scene used Telegram for communicating radical views. The dominant vehicle of disinformation was alternative websites whose content projected into disinformation groups on social media [13, p. 15–16].

As the recent analyses show, throughout 2021, at the 46 active pro-Kremlin websites in Czechia more than 197,000 articles were published and disinformation trends were strengthening. The three biggest Czech disinformation websites in Czechia have had an average of almost 14,500,000 user visits per month. As disinformation expert Miloš Gregor concludes: 'the published data confirm the fact that Russia was preparing itself for the war for a long time. It is exactly Russia who mostly supports the disinformation and propaganda websites in Czech Republic' [40].

The most positive outcome of our analysis might be the fact that the new government showed after 24 February 2022, and beforehand with its Policy Statement, that it considers hybrid threats, including disinformation campaigns, as activities contrary to the principles of democracy and as a very serious threat to security. The extraordinary blockade of selected disinformation websites showed that the government, and even more the intelligence services, continually collect the necessary data and information on subversive actors. Furthermore, the new government also showed its interest in the work of the intelligence services and gave clear support for these activities. Fendrych points out the symbolic arrangement of the press conference, where the director of BIS, Michal Koudelka, presented the recent Annual report with Prime Minister Petr Fiala seated next to him. As Koudelka recalled: 'It is for the first time after 24 years, when the representative of the state comes forward before the journalists after visiting BIS. Last time it was President Václav Havel'. Koudelka also noted that 'we (BIS – quoted by Ladislav Cabada) understand disinformation as one of the biggest recent risks for the security of the Czech Republic'. He considers the following to be extremely dangerous: 'state actors who attack our nation by disinformation. The state actor number one is Russia, followed by China, but also Iran is here, and others' [39].

To sum up, during the year 2022 the new Czech government overcame the previous period characterised by questioning the activities of the intelligence services, this period was represented mainly by President Zeman. The new Czech government also returned officially and clearly to the necessity of considering disinformation campaigns as an important part of a new hybrid threat to the nation. Of specific importance is the creation of a new institutional framework, rooted in legislation in the struggle against hybrid threats. Only such a system will be robust and stable enough to endure future attacks from both outside and more importantly from internal actors who are a part of the disinformation scene and/or are using disinformation strategies [41].

Funding

This article is the result of Metropolitan University Prague research project no. E46–66 based on a grant from the Internal Grant System.

References

[1] Š. Waisová, "Central Europe in the New Millenium: The new great game? US, Russian and Chinese interests and activities in Czechia, Hungary, Poland and Slovakia," *Revista UNISCI*, no. 54, pp. 29–48, 2020.

[2] M. Bokša. (2019, June 15). *Russian Information Warfare in Central and Eastern Europe: Strategies, Impact, Countermeasures, German Marshall Fund.* [Online]. Available: https://www.jstor.org/stable/resrep21238. [Accessed: Oct. 24, 2022].

[3] L. Cabada, "Changes in Mutual Relations between Czech Social Democrats and Communists after 2000 and Strengthening of Anti-Communism in Czech Society and Politics," *Politické vedy*, vol. 18, no. 4, pp. 8–31, 2015.

[4] L. Kopeček, P. Pšeja, "ČSSD a KSČM: na cestě ke spojenectví?," *Politologická revue*, vol. 13, no. 2, pp. 35–59, 2007.

[5] P. Popálený, L. Cabada, "100 Years of the Communist Party in the Czech Lands: A Comparison of the Inter-war and Post-Transitional Situation," *Journal of Comparative Politics*, vol. 15, no. 1, pp. 4–23, 2022.

[6] L. Cabada, "Nova personalizirana stranačka politika v Češkoj," *Političke analize*, vol. 9, no. 33–34, 2018.

[7] L. Cabada, "The Human Rights Aspect of Czech Foreign Policy – A Change of Course after 2013," in *Removing Barriers, Promoting Responsibility The Czech Centre-Right's Solutions to the Political Challenges of 2021*, L. Tungul, Ed. Prague: Wilfred Martens Centre for European Studies, Konrad-Adenauer-Stiftung and TOPAZ, 2020, pp. 12–21.

[8] P. Kratochvíl, "Von Falken und Russlandsfreunden. Die tschechische Debatte über die EU-Sanktionen," *Osteuropa*, vol. 64, no. 9–10, pp. 67–78, 2014.

[9] J. Hrdlička. (2022, Oct. 28). *Zeman a ´čučkař´ Koudelka. Stojí za tím osobní msta prezidenta za odposlechy BIS, Echo24.cz.* [Online]. Available: https://echo24.cz/a/ScwJE/zparvy-do-mov-zeman-cuckar-koudelka-msta-odposlechy-bis. [Accessed: Oct. 29, 2022].

[10] J. Dvořáková, T. Syrovátka. (2021). *Vrbětice: Case Study of Czech Resilience against Hostile Propaganda, Prague Security Studies Institute (PSSI) Perspectives.* [Online]. Available: https://www.pssi.cz/download/docs/8744_pssi-perspectives-12-vrbetice-case-study.pdf. [Accessed: Oct. 21, 2022].

[11] I. Gaba. (2022, Mar. 22). *Miloš Zeman sloužil ruské propaganda. Data jeho ´zmýlení v*

Putinovi ´ vyvracejí, HlídacíPes.org. [Online]. Available: https://hlidacipes.org/ivan-gabal-milos-zeman-slouzil-ruske-propagande-data-jeho-zmyleni-v-putinovi-vyvraceji/. [Accessed: Oct. 29, 2022].

[12] L. Spalová, *Media – Migration – Politics. Discursive Strategies in the Current Czech and Slovak Context*. Berlin: Peter Lang, 2022.

[13] Security Information Service. (2022). *Annual Report 2021.* [Online]. Available: https://www.bis.cz/annual-reports/annual-report-of-the-security-information-service-for-2021-566c4b7f.html. [Accessed: Oct. 19, 2022].

[14] M. Mareš, J. Kraus, J. Drmola, "Conceptualisation of Hybrid Interference in Czechia: How to Make it a Practically Researchable Phenomenon?," *Politics in Central Europe*, vol. 18, no. 3, pp. 343–354, 2022.

[15] J. Daniel, J. Eberle, "Hybrid Warriors: Transforming Czech Security through the 'Russian Hybrid Warfare' Assemblage," *Sociologický časopis / Czech Sociological Review*, vol. 54, no. 6, pp. 907–932, 2018.

[16] A. Tucker, *Democracy Against Liberalism. Its Rise and Fall*. Cambridge and Medford: Polity Press, 2020.

[17] P. Norris and R. *Inglehart, Cultural Backlash: Trump, Brexit, and Authoritarian Populism*. London: Cambridge University Press, 2019.

[18] A. Ágh, *Declining Democracy in East-Central Europe. The Divide in the EU and the Emerging Hard Populism*. Cheltenham: Edward Elgar Publishing, 2019.

[19] L. Cianetti, J. Dawson, S. Hanley, Eds., *Rethinking ´Democratic Backsliding´ in Central and Eastern Europe*. London and New York: Routledge, 2019.

[20] H. Allcott, M. Gentzkow. (2017). "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, doi: 10.1257/jep.31.2.211.

[21] J. Syrovátka, J. Hroch. (2018). *Czech Election in an Era of Disinformation, Prague Security Studies Institute (PSSI)*. [Online]. Available: https://www.pssi.cz/download//docs/8404_545-presidential-election-2018-analysis.pdf. [Accessed: Oct. 21, 2022].

[22] J. Baqués-Quesada, G. Colom-Piella, "Russian Influence in Czechia as a Grey Zone Casy Study," *Politics in Central Europe*, vol. 17, no. 1, pp. 29–56, 2021.

[23] L. Valášek, J. Horák. (Sep. 27, 2021). *Babiš zdržuje boj státu proti desinformacím. Místo jejich řešení zdůrazňuje migraci, Aktualne.cz.* [Online]. Available: https://zpravy.aktualne.cz/domaci/experti-varuji-ze-cesko-je-proti-dezinformacim-bezbranne-bab/r~d1caf58c1b8211ecbc3f0cc47ab5f122/. [Accessed: Oct. 19, 2022].

[24] R. Bartoníček. (2022, Feb. 18). *Žádná cenzura, ujišťuje Fialova vláda. Boj s desinformacemi vykopne děkovací kampaní, Aktualne.cz.* [Online]. Available: https://zpravy.aktualne.cz/domaci/vlada-boj-s-dezinformacemi/r~08f399ca8efa11ec8d900cc47ab5f122/. [Accessed: Oct. 17, 2022].

[25] J. Jetmar. (2022, Feb. 10). *Václav Moravec představil středoevropský projekt pro boj s dezinformacemi, Mediar.cz.* [Online]. Available: https://www.mediar.cz/vaclav-moravec-predstavil-stredoevropsky-projekt-pro-boj-s-dezinformacemi/. [Accessed: Oct. 29, 2022].

[26] *Media Literacy Observatory for Active Citizenship and Sustainable Democracy.* [Online]. Available: https://www.interreg-danube.eu/approved-projects/melia-observatory. [Accessed: Oct. 29, 2022].

[27] M. Gregor, P. Mlejnková, "Facing Disinformation: Narratives and Manipulative Techniques Deployed in Czechia," *Politics in Central Europe*, vol. 17, no. 3, pp. 541–564, 2021.

[28] L.S. Bánkuty-Balogh, "Novet Technologies and Geopolitical Strategies: Disinformation Narratives in the Countries of the Visegrád Group," *Politics in Central Europe*, vol. 17, no. 2, pp. 165–195, 2021.

[29] J. Soukup, A. Trousilová. (2022, Oct. 28). *Proti vládě demonstrovaly desetitisíce lidí. Převezmeme tuto zemi, hlásil organizátor, Novinky.cz.* [Online]. Available: https://www.novinky.cz/clanek/

domaci-protivladni-demonstrace-na-vaclavskem-namesti-40412898. [Accessed: Oct. 29, 2022].

[30] Government of Czechia. (2022, Jan. 6). *Policy Statement of the Government.* [Online]. Available: https://www.vlada.cz/en/jednani-vlady/policy-statement/policy-statement-of-the-government-193762/. [Accessed: Oct. 19, 2022].

[31] J. Horák. (2021, Feb. 3). *Možný průlom v boji proti hybridním hrozbám, vládní tým vede Babišův vojenský poradce, Aktualne.cz.* [Online]. Available: https://zpravy.aktualne.cz/domaci/hybridni-hrozby-brs-matous/r~f04af6145bc611eb8b230cc47ab5f122/. [Accessed: Oct. 29, 2022].

[32] Echo.24. (2022, Feb. 25). *Česká vláda nechala zablokovat některé weby, které označila jako dezinformační.* [Online]. Available: https://echo24.cz/a/ScjgR/ceska-vlada-nechala-zablokovat-nektere-weby-ktere-oznacila-jako-dezinformacni. [Accessed: Oct. 24, 2022].

[33] Ministry of Internal Affairs, *Government of Czechia, Trestněprávní úprava.* [Online]. Available: https://www.mvcr.cz/chh/clanek/dezinformacni-kampane-trestnepravni-uprava-trestne-pravni-uprava.aspx. [Accessed: Oct. 19, 2022].

[34] P. Prchal. (2022, Sep. 23). *K soukromoprávním hlediskům blokace dezinformačních webů (studie), Advokátní deník.* [Online]. Available: https://advokatnidenik.cz/2022/09/23/k-soukromopravnim-hlediskum-blokace-dezinformacnich-webu-studie/. [Accessed: Oct. 17, 2022].

[35] T. Hübscherová. (2022, Oct. 19). *Rusko propaganda u nás šíří nejvíce sami Češi. Je třeba hledat příčinu dezinformací, varuje expert, Forum 24.* [Online]. Available: https://www.forum24.cz/ruskou-propagandu-u-nas-siri-nejvice-cesi-je-treba-hledat-pricinu-dezinformaci-tvrdi-expert/?utm_source=www.seznam.cz&utm_medium=sekce-z-internetu#dop_ab_variant=0&dop_source_zone_name=hpfeed.sznhp.box. [Accessed: Oct. 19, 2022].

[36] M. Fišer. (2022, May 25) *Dezinformační weby v Česku opět fungují. Až na Aeronet, Novinky.cz.* [Online]. Available: https://www.novinky.cz/clanek/internet-a-pc-dezinformacni-weby-v-cesku-opet-funguji-az-na-aeronet-40398111. [Accessed: Oct. 23, 2022].

[37] J. Cibulka. (2022, Apr. 18). *Zákon nevzniká, analýzy nejsou. Legislativa proti dezinfu bude mít minimální dopad, varují právníci, iRozhlas.* [Online]. Available: https://www.irozhlas.cz/zpravy-domov/dezinformace-blokovani-cenzura-zakon_2204180500_cib. [Accessed: Oct. 21, 2022].

[38] ČTK. (2022, Oct. 26). *Soud rozhodl o Tušlovi s Čermákem. Dezinformátoři půjdou za mříže, Czech Press Agency.* [Online]. Available: https://www.forum24.cz/v-kladne-zacal-soud-s-muzi-obzalovanymi-za-nenavistne-vyroky-o-ukrajincich/. [Accessed: Oct. 29, 2022].

[39] M. Fendrych. (2022, Oct. 19). *Fiala podpořil Koudelku. Snad bude odteď dbát na bezpečnost podobně jako na ekonomiku, Aktualne.cz.* [Online]. Available: https://nazory.aktualne.cz/komentare/bude-konecne-bezpecnost-stejne-dulezita-jako-ekonomika/r~49b991de4ec-c11edbe29ac1f6b220ee8/?utm_source=www.seznam.cz&utm_medium=sekce-z-internetu#dop_ab_variant=0&dop_source_zone_name=hpfeed.sznhp.box. [Accessed: Oct. 21, 2022].

[40] B. Novotná, P. Juna. (2022, Oct. 3). *Češi si proti desinformacím nevěří. Pomoct může ´očkování´, seznamzpravy.cz.* [Online]. Available: https://www.seznamzpravy.cz/clanek/fakta-dezinformace-jsou-na-vzestupu-a-cesi-si-v-jejich-odhalovani-neveri-215774. [Accessed: Oct. 17, 2022].

[41] D. Rychnovská, M. Kohút, "The Battle for Truth: Mapping the Network of Information War Experts in Czechia," *The New Perspectives*, vol. 26, no. 3, 57–87, 2018.

How Are Czech Individuals Willing to Protect Themselves: A Comparison of Cyber and Physical Realms

Jan Kleiner | **Department of Political Science, Faculty of Social Studies, Masaryk University, Brno, Czech Republic**, ORCID: 0000-0001-9950-410X

Jakub Drmola | **Department of Political Science, Faculty of Social Studies, Masaryk University, Brno, Czech Republic**, ORCID: 0000-0003-4275-2115

Miroslav Mares | **Department of Political Science, Faculty of Social Studies, Masaryk University, Brno, Czech Republic**, ORCID: 0000-0002-7102-3205

Abstract

Endpoint users are usually viewed as the highest-risk element in the field of cybersecurity. At the same time, they need to be protected not just from the individual-level prism but also, from the state's perspective, to counter threats like botnets that harvest weakly secured endpoints and forge an army of so-called zombies that are often used to attack critical infrastructure or other systems vital to the state. Measures aimed at citizens like the Israeli hotline for cybersecurity incidents or Estonian educational efforts have already started to be implemented. However, little effort is made to understand the recipients of such measures. Our study uses the survey method to partly fill this gap and investigate how endpoint users (citizens) are willing to protect themselves against cyber threats. To make results more valid, a unique comparison was made between cyber threats and physical threats according to the impact which they had. The results show statistically significant differences between comparable cyber-physical pairs indicating that a large portion of the sample was not able to assess the threat environment appropriately and that state intervention with fitting countermeasures is required. The resultant matrix containing frequencies of answers denotes what portion of respondents are willing to invest a certain amount of time and money into countering given threats, this enables the possible identification of weak points where state investment is needed most.

Corresponding author:
Miroslav Mares, Department of Political Science, Faculty of Social Studies, Masaryk University, Brno, Czech Republic;
ORCID: 0000-0002-7102-3205;
mmares@fss.muni.cz

Keywords
cyber-physical comparison, cybersecurity, state-endpoint user relationship,

Cite this article as: J. Kleiner, J. Drmola, M. Mares, "Protect Themselves: A Comparison of Cyber and Physical Realm," ACIG, vol. 1, no. 1, pp. 235–252, 2022,
DOI: 10.5604/01.3001.0016.1322

1. Introduction

The realm of cybersecurity does not only consist of data, computers, routers, and other devices. An endpoint user is one of the basic elements, layers more precisely, as well. A cyberattack typically needs an attack vector, a way to compromise a target system. In most cases, it is the very human error (e.g. incited by phishing) that enables a breach. A user is thus perceived as the weakest link in the cyber structure [1, 2]. Therefore, it is only logical that a specialized branch of cybersecurity research is devoted to the understanding of human roles and behaviour in cyberspace. This quickly growing branch is called cybersecurity behavioural research, it consists of more than five hundred academic publications [2]. This paper aims to be one of them as it investigates the willingness of endpoint users to protect themselves based on the possible impact of various cyber threats.

However large and significant the cybersecurity field has become over recent decades, the general public, including public officials, can often find it challenging to navigate and understand. Therefore, this study comes with a standardized comparison with the physical world and its threats that are more understandable for individuals in our sample. The comparison represents a unique attempt. Hence, robust methodological sections are exhaustively laid out later in the study so the research can be easily replicated. It can also provide a useful benchmark for both readers of this study and participants in the survey described below. This feature creates an obvious obstacle in any case. How to compare such different realms?

The main research question is: how are Czech endpoint users willing to protect themselves against cyber threats and physical threats? As is later explained, the whole study leans towards inductivist logic; hence the secondary goals include a thorough investigation of relationships between examined variables and a detailed description of the methodological process. Emphasis is laid here on the cyber-physical comparison in order to provide a functional framework for possible replication. Given the emerging implementation of measures by a state aimed at securing its citizens (discussed in the next section) another inquiry arises as well: are such measures justifiable by a clear need of citizens?

1.1. The State and its Citizens in Cyberspace: A Need for Research

The scope and motivation of this study significantly overlap with the concept of "secured citizens, secured state". When present in vast numbers insufficiently secured users and their devices can pose a threat to public

administration and critical infrastructure in the form of pivoting attacks or the bring-your-own-device (BYOD) phenomenon. States have hence begun to bring this topic to public debate in various forms and intensities. For example, from January 2020 California banned the usage of default and hard-coded passwords in all devices that are to be sold in this Silicon Valley state [3]. Weakly secured electronic devices are widely used in botnet attacks as well as other forms of attack. The aforementioned law is thus aimed at reducing this type of threat [4]. A less coercive tool was created in Israel – a hotline that businesses and private individuals can contact if they suspect they are victims of a cyberattack [5]. Both cases illustrate an endeavour from the state to more or less incorporate its citizens into cybersecurity processes.

It has already been mentioned that quite a lot of research is devoted to human behaviour connected with cyberspace and cybersecurity. Inadequate academic attention is, however, dedicated to understanding the willingness of users to protect themselves against various online threats. Such an understanding could serve as a knowledge base for public authorities to design effective measures to protect their citizens. Efforts like the Californian law are emerging to protect citizens in cyberspace, and consequently to protect the state itself, but recipients of such measures have not been adequately understood. This is an application of imperfect means utterly preventable with proper research. A better understanding of the endpoint user's willingness to protect him or herself could increase the efficiency of the whole process as a state can invest in areas where users are not willing to invest themselves and *vice versa*. Another purpose of this paper is to open a discussion on this topic and to provide a stepping stone for further research.

1.2. Relevant Academic Context

Even though cybersecurity behavioural research is quite a large field with hundreds of academic papers [2], the comparison between cyber and physical threats or dimensions is unique. This is especially so in the prism of the state actor and security research. We employed a relatively structured approach in identifying the relevant literature. The SCOPUS database was searched using the keywords "willingness" and "cybersecurity", resulting in 42 papers and 112 articles for the search string "cyber AND comparison AND threat AND physical OR real". All abstracts were then close-read and assessed for their relevance to our research.

Most researchers devote their attention to either of these two realms, i.e. [6, 7]. The latter paper measured the willingness of individuals to pay and adopt cybersecurity training on a Swedish-based sample and how this was affected by worry about various cyber threats. They found no correlation whatsoever between these two variables.

Thematically close is the paper authored by Furman et al. [8], which examines users' perception and knowledge of cyber threats through an interview conducted with 40 participants (most of them with college education). The paper states that most participants rely on third parties to provide them with online security (e.g. state, software companies, and

banks). A large portion (44%) of them are not able to generally define the most common cyber threats, such as phishing, keylogger, botnet or spyware [8]. Both these insights are valuable for us as the first one justifies an increase in the state's interest in the cybersecurity of its citizens, and the latter supports the usefulness of comparison between cyber and physical realms.

Bauman & Newman's [9] study lays down a foundation for our cyber and physical domain comparison. In this study the authors juxtapose cyberbullying and its classical form. The sample of 588 students was presented with a survey measuring the perception of manifestations of both types of bullying. In other words, the impact of various threats were used to measure very incomparable domains. This is a crucial intake of Bauman & Newman's article [9] as a similar approach is used in this study.

Similar to our conceptualization of the willingness of endpoint users to protect themselves is a study by Fagan & Khan [10] on perception in a user's decision-making process concerning the costs and benefits calculation. Their article examines the motivation of users to ignore cybersecurity advice. Assuming a user's rationality, they presume that a user's decision is the result of the endeavour to maximize benefits and minimize costs. After analyzing 12 combinations of motivators of benefits, risks, and costs along with the individual and societal dimensions, they conclude that a benefit is a crucial motivator if the user associates that benefit with an action [10].

2. Methods and limits

The data in this study mainly comes from an online survey which was conducted in the Czech Republic. There are limitations inherent to this source, these methods and to the tools used, they are introduced to meet the transparency criterion of the scientific method. The general approach to the survey-creating process was taken from an article by Schaeffer & Presser [11] which lists the possible, mostly respondent-related errors and cognitive biases as well as how to avoid them. In addition, the survey-creating process also incorporates the three-rule approach from Bourke, Kirby and Doran [12] which states that a respondent must both understand the question and be willing and capable of answering it.

The sample that came out of the survey consists of 247 participants, 96 (38.9%) men and 151 (61.1%) women. As regards education, most of the participants (51.4%) had a college degree (one had another form of higher education), 26.7% had a high school education, and 21.9% had only primary education. As the survey works closely with the economic situation of participants, it is essential to mention that 66.4% were students, and 31.5% were economically active (employed or self-employed).

The survey's intersections of axes X and Y resulted in 36 main-body questions which produced 90 values per matrix (Tab. 3.). These values provided information about the willingness of endpoint users to protect themselves from various impact intensities. Five questions were used to measure the basic cybersecurity knowledge of respondents, this provided a cybersecurity score variable (max. value was 7, median 5, and mean 5.24) and another five questions measured socioeconomic status

based on monthly income. The mean value of the income variable was 23 730 CZK (approx. 860 EUR), and the median was 15 000 CZK (approx. 545 EUR). Although approx. 25% of respondents refused to share their monthly income. The rest of the variables measured willingness to protect themselves against various cyber and physical threats.

Such a sample is not representative, i.e. we cannot infer conclusions based on the sample data to the whole intended population. This is the most substantial limit of the study. Nevertheless, let us reiterate its purpose, which is to provide a premier comparison and insight into developing a state-citizen relationship in cyberspace research. Hence, if viewed through an inductivist prism, the study can still bring valuable benefits to the field of cybersecurity behavioural research, despite the unrepresentative sample.

2.1. A Comparison of Cyber and Physical Domains

One of the key features in the survey-making process is standardization, this increases reliability. Every respondent has to receive the same set of questions, which have to be processed and analyzed the same way throughout the process, so that interviewer error is minimized [13]. The same principle logically applies to the comparison of two sets of questions (cyber and physical). But how to transpose theory into this study's practice? Two hindrances must be overcome.

In section 1.2 the Bauman & Newman [9] article was briefly introduced, it offers a potential solution to the comparison question. The perception of cyber and classical bullying was measured by the impact of their manifestations which provides a rather useful and straightforward approach for comparison of phenomena which are different. The method of Bauman & Newman [9] is designed for concrete threats and their manifestations. However, the aim of our study is more general as it takes into consideration all possible threats. Therefore, the method must be altered to fit here. That brings us to the second hindrance to be solved for the comparison to work correctly.

There is no reasonable way to incorporate every single cyber and physical threat and to compare all of them in a standardized manner to achieve the desired results. One of the authors faced a similar issue with the quantitative risk assessment of eGovernment in the Czech Republic [14]. Inspired by the Czech National Cyber and Information Security Agency (NÚKIB, former National Centre for Cyber Security) a matrix and interval approach was implemented based upon their conduct in such analysis. Instead of a particular description of each of the countless threats that would later enter an analysis, a broader interval form allowed us to classify them into categories which enter the analysis as entities in manageable quantity.

To summarise, this study uses the approach of Bauman & Newman [9] to make the difficult comparison between physical and cyber threats through their impacts, with intervals to reduce numerous cyber and physical threats into categories. The matrix then serves as a way to arrange the data for comparison.

2.1.2. Conceptualization of key terms

Now that the central assumptions are set and clear, we can move to the conceptualization of key phenomena. Zeman defines a threat as a "primary, independently existing phenomenon, which can or wants to damage a concrete value"[15]. A dualist division of intentional and unintentional threats stems from this definition. To reasonably reduce the research subject, only intentional (e.g. perpetrated by a human being) threats are taken into consideration here.

A cyber threat is simply defined as "anything that has the potential to cause serious harm to a computer system" [16]. A "physical" threat does not represent a potent term in the field of political science. However, it plays a role in jurisprudence. In order not to encroach on the field of law, as this paper concerns political science, the study combines Zeman's [15] definition of threat with the concept of "physical", which is "anything that has a material existence: perceptible especially through the senses and subject to the laws of nature" [17].

The willingness of endpoint users to protect themselves is not a generally used or accepted term either. Hence, it needs to be conceptualized indirectly, as well. Zeman's [15] equation of risk (risk = the multiplication of threat, vulnerability, and value divided by countermeasures) can be helpful here. The willingness to protect oneself is *de facto* the willingness to apply countermeasures. While assuming rationality, it is a calculation of costs and benefits. The latter represents here the absence of damaging impacts. The willingness to protect oneself can thus be perceived as a propensity to invest in something trying to prevent a threat from happening. There are several forces in place that largely influence the essence of that "something". The survey must be kept brief to increase the response rate. It must also be comparable across both cyber and physical realms and easily imaginable for respondents. Hence, the study works with two concepts of investment: financial and time. To anchor it more in theory, Zeman [15] interlinks financial investment with countermeasures as well. The time component was added to reflect the nature of the cyber domain better as there are a vast number of countermeasures that require time rather than a direct financial investment, such as the invention of strong passwords or learning how to conduct cyber hygiene properly.

Similarly, complicated conceptualization is tied to impacts. This study takes inspiration from the approach of the NÚKIB, which lists threat impacts on life and health, economics and finance, reputation and the upholding of laws and regulations [14]. Unlike NÚKIB, this analysis focuses on an individual level, not on the state level of critical infrastructure. The cyber-physical comparison must be kept in mind as well. Therefore, the common intersection here takes the form of three categories of impact on an individual's life and health, economics and data.

2.1.3. Matrices and operationalization

To elaborate more on the research question, the primary goal of the study was to measure the willingness of endpoint users to protect themselves

(a dependent variable) based on their perception of threats represented by their impacts (an independent variable). Hence, a matrix consisting of two axes, one for each variable, is a fitting and clear way to structure the data. As there is a need for the comparison of physical and cyber domains in a standardized way, two standardized matrices are used.

Variables need to be operationalized for the measurement to be possible. Cox suggests that a respondent should be offered five to nine options to choose from [18]. Nevertheless, there is another caveat in place stating that the longer the survey, the lesser the response rate. Hence, five categories of both time and financial investments are offered in survey questions measuring the willingness of the endpoint users. This variable lies on axis X (see Tab. 1.).

The investment categories need to be exhaustive so that each respondent can fit in one of them. Here are the intervals. Time investments are equally divided as follows:

- 0 minutes (no investment at all)
- <1min.; 59min.>
- <1hr.; 24hrs.>
- <24hrs.; 7 days>
- 7 days or more denoted as 'several weeks'

The financial investments are based on the median pay in the Czech Republic, which is 27 600[1] CZK [19], so the respondents can easily relate to the given answers. The stratification of values was adjusted due to feedback from the focus group pilot testing:

- 0 CZK (no investment at all)
- <1 CZK; 999 CZK>
- <1k CZK; 9 999 CZK>
- <10k CZK; 27 599 CZK>
- 27 600 CZK or more

1 —— Approx. 1000 EUR.

Table 1. Axis X of the matrix.

Willingness of endpoint users to protect themselves (axis X)									
Time investment					Financial investment (in CZK) 1 EUR = approx. 27.5 CZK				
0 min.	59 min.	24 hrs.	7 days	Several weeks	0	1–999	1k-9 999	10k – 27 599	27 600+
T1	T2	T3	T4	T5	F1	F2	F3	F4	F5

The axis Y of the matrix is left for the independent variable, the impacts of cyber-physical threats. In an operationalized form, it is more specific to refer to them as intensities of impacts. There are three degrees of intensity used to provide a respondent with two extreme and one middle option. If only one more category had been added, the number of questions would have increased by 30 per matrix, which would have significantly threatened the response rate.

As this study is limited to only intentional, hence usually criminal threats, it relies on the Czech Criminal Code [20]. Paragraph 122 defines "serious injury" as "mutilation, loss or substantial reduction in fitness, limb paralysis, loss or substantial impairment of sensory function, damage to an important organ, disfigurement, induction of abortion or killing of the foetus, torturous distress, or prolonged impairment of health" [20]. A time of convalescence is used in the given questions in order to present a relatable and imaginable form to a respondent. At the same time, the 'high intensity' option mentions the possibility of death as the ultimate form of an impact. The other two intensities are equally distributed: low <1, 7 days>, medium <8, 27 days> (plus high intensity: 28 days or more plus the possibility of death) (Tab. 2.). The number of days was chosen due to the hypothesis that the numerical form of the operationalized variable reduced the risk of misinterpretation [11], the figures were arrived at in consultation with physicians.

Economic impacts rely heavily on the personal situation of the given respondent. In order to be relatable to as many respondents as possible, the same approach as with the financial investments above is used. Based on the median for pay, which is approx. 27 600 CZK [19], the intervals of three intensities are equally distributed (Tab. 2.).

Regarding the data impact category, the 'CIA triad' is used for operationalization. This information security concept can be applied to both the cyber and physical worlds. It stands for Confidentiality, Integrity and Availability of data, it perceives these factors as the desirable and protected values of a reference object, namely data [21]. The CIA concept can apply to a physical world as well. Confidentiality can be compromised by ID card theft, availability by their loss and integrity by the shredding of important documents. The distribution of intensities is intuitive here. High intensity represents the disruption of all three attributes, namely theft and complete and permanent loss of data in any form. Medium intensity is the complete and permanent loss of data (without theft), and low intensity takes the form of a temporary denial of accessibility.

Table 2. Axis Y of the matrix.

The intensity of impact (axis Y)	Life and health (convalescence time in days)	Low (LH1) Medium (LH2) High (LH3)	1–7 8–27 28+ or death
	Economic impact (in CZK; 1 EUR = approx. 27.5 CZK)	Low (E1) Medium (E2) High (E3)	1 – 13 799 13 800 – 27 599 27 600+
	Data (CIA triad damage)	Low (D1) Medium (D2) High (D3)	Temporary denial of accessibility Complete and permanent loss of data Complete and permanent loss of data by theft

2.1.4. Collection of survey data

The intersections of both axes determined the form of questions in the survey. Each category for intensity of impact contained an example of both a cyber threat and a physical threat, this provided as clear and relatable a set of questions as possible for each respondent, thus increasing the validity of the survey. Therefore, the resulting survey consisted of 18 questions for each matrix measuring the influence of the aforementioned impacts on the willingness of endpoint users to protect themselves. A respondent was offered five options denoting how much they were willing to invest, in either time or money, to mitigate the risk. At the very beginning of the survey respondents were briefed that the more they invested, the more the risk decreased. Although an oversimplification, this narrative of linearity was intended to make answering the questions easier while measuring the influence of intensity of impacts on willingness.

Each of the 36 main-body questions contained possible investments denoted by precise numerical intervals as well as examples of countermeasures that corresponded with given intervals of time and financial investment, e.g. "an online cybersecurity course from 24 hours inclusive to 7 days exclusive". The first draft of the survey contained only numerical expressions, but specific examples were added due to the results of the focus-group[2] pilot testing. The focus-group feedback also mentioned the need for information, especially about cybersecurity and cyber threats, so that participants could make an informed decision upon which type of investment to choose. This could damage the ecological validity, but the focus group's demand was absolute, so we chose introductions which were as brief as possible to meet this demand whilst minimizing the damage.

2 —— The focus group consisted of 20 participants with various ages, educational levels, incomes and cybersecurity knowledge.

The survey also contained five questions on socio-economic standing and five measuring the level of cybersecurity knowledge; the latter stem from Google [22] and UC Berkeley [23] basic security tips. Both entities recommend strong password policies and frequent software updates. Furthermore, they warn about identity theft, spoofing and phishing attacks and recommend a cautious attitude when working with suspicious emails which request sensitive data and access data. Verifying URLs and the need for data backup are mentioned by Google and UC Berkeley as well [22, 23]. The socioeconomic section inquired about a respondent's age, education, type of employment (tailored for the Czech market), and income. While the first four questions were compulsory, the last one was optional due to its intrusiveness.

The survey was distributed via various social media groups; therefore, the sample is convenient and not representative of the Czech population. However, as the study's logic is qualitative, we were aiming for theoretical saturation rather than costly statistical representativeness, which we achieved as shown in Tab. 3.

3. Data analysis

The results were statistically analyzed using IBM's SPSS software ver. 25. The assessment of the data centrality (means and medians) for respondents

is laid down in section 2.1, describing the sample. Despite histograms with promising skewness and kurtosis, none of the variables had normal distribution according to the Kolgomorov-Smirnov test, which allows for only robust further statistical testing.

3.1. Basic descriptives

We started the analysis with frequencies (Tab. 3.) and by looking for relationships in socioeconomic and cybersecurity-knowledge variables. In most cases relationships were either not statistically or factually significant. Only gender correlated weakly (0.187) with cybersecurity score ($p < .01$). Crosstab results then revealed more men in the higher echelons (cybersecurity score 6 and 7) than women, meaning that men tend to be slightly more knowledgeable and cautious in cyberspace than women.

Crosstabs were also used for examination of the willingness of individuals to protect themselves. The risk of cyber threats having an impact on human life or health was shallow [24]. Despite the real-world probabilities, respondents were eager to put the biggest time and financial investments in countermeasures against the life or health threats compared to the other two categories (economic and data) of impacts. That is, nevertheless, only a logical and anticipated conclusion. On the other hand, a threat to life and health can more easily and probably occur in the physical realm. The highest forms of investments were, in the physical matrix, concentrated more on the economic impacts than in the life or health categories. This suggests that respondents were more afraid of fraud and theft than violent crime. That would be a surprising finding if not for the fact that the Czech Republic is one of the safest countries in the world [25].

A more interesting observation appears if we compare time and financial investments for each of the intensities of impact then graphically differentiate that which has the higher frequency of answers (Tab. 3.). This produces something called “the breaking interval” which denotes a threshold after which individuals, on average, are more likely to invest time than money or *vice versa*. The breaking interval, e.g. low intensity of impact on life and health in the cyber aspect of the matrix, is <1 CZK; 999 CZK>. Higher investments, in this case, are preferable in the form of spending time rather than money. If the intensity in the same impact category rises to the medium level, the breaking interval shifts to <1 000 CZK; 9 999 CZK>. The identification of the breaking interval could, for example, serve for the creation of tailor-made cybersecurity education, subsidized anti-virus software subscription or other measures made by a state for its citizens.

3.2. Cyber-physical comparison

The frequencies contained in Tab. 3. are self-explanatory, so let us move to the trickier comparison of the cyber and physical worlds. As none of the survey variables is normally distributed, non-parametric testing had to be done. Using the Wilcoxon signed-rank test, we compared pairs (one for cyber and one for the physical realm) for a given investment type

intersected with the intensity of the given impact. The test then returned significance and the number of positive ranks (number of respondents willing to invest more) and negative ranks (number of respondents willing to invest less) for each pair.

All of the pair comparisons were significant (Tab. 4.), except for the willingness to invest time against cyber and physical threats with low and high intensities of impact on human life and health (hence, these two were crossed out from the table). Before the results are examined, it is essential to reiterate that respondents scored relatively highly in the basic cybersecurity knowledge test (mean 5.24; median 5; max. value 7). Could this have had any significant impact on the results?

We argue that it very likely could not. Using the Mann-Whitney test, which examines the differences between two conditions and between two different groups (this is called 'the grouping variable'), we found that none of the variables we used for grouping – cybersecurity score, age, income and education in their recoded dichotomous form – had any significant effect on the two test variables of the given willingness and impact intensity intersections for either the cyber or physical realm (e.g. willingness to invest zero time to prevent the low intensity of the impact of a cyber or physical threat). In other words, the level of cybersecurity knowledge or age of a respondent could not explain their willingness to invest more or less in comparable cyber and physical countermeasures.

The results turned out to be quite predictable in the 'life and health' category because the general trend reflected the reality quite well, especially in terms of probabilities of threats occurring. This is interesting as the portion of respondents who performed a risk assessment, meaning they did not choose the same investment for both of the realms (Tab. 4.), performed, on average, the assessment well. Life and health threats are rare in the cybersecurity field, unlike in the physical realm [24]. In accordance with this, 87 to 99 respondents (depending on the type of investment and the intensity of impact) were willing to invest more in physical countermeasures. However, the differences between well-assessing and badly-assessing (positive minus negative ranks) individuals were not substantial. Also the numbers of ties were high (around 100), indicating that large portions of people did not assess or properly distinguish between cyber and physical threats. A point of note is that the number of ties is very similar across the comparisons showing the consistency we elaborate on in the discussion.

More mixed and ambiguous trends occurred in the economic category. As regards time investments, there were more negative than positive ranks indicating that more people had been willing to invest more in measures countering cyber threats (the survey mentioned ransomware) than in measures countering physical ones (theft, fraud, and embezzlement). By using worldwide statistics as well as those from the Czech police, Kleiner (2020) argues that in the Czech Republic there are higher frequencies and more severe damage on the side of the mentioned physical threats [24]. In order to cope more with reality, the reported trend should thus have an opposite direction which can be found in financial investments (more positive than negative ranks).

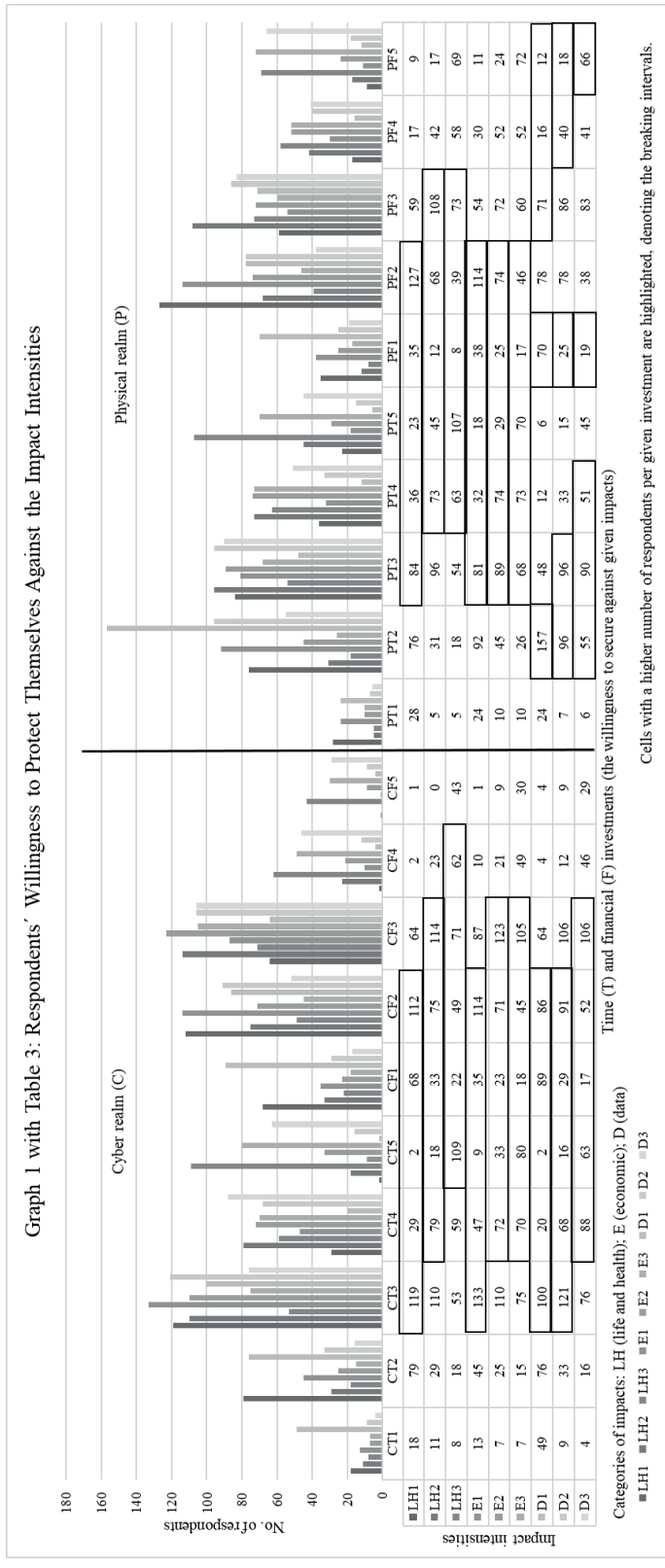
As the life and health tier was chosen to better suit the physical world, the data one was intended dominantly for the cyber realm. The

survey mainly emphasized email, internet banking, identity theft, and similarly ID cards, passports, sensitive documents, stalking and other forms of surveillance and intrusion. Trends here are very similar to the economic impacts category. As the Czech police statistics are not sufficiently detailed, it is hard to conclude if the trend is following reality [26]. Cyberattacks aimed at obtaining personal data are rampant and extensive on a global scale [27]. We should, therefore, see much more negative than positive ranks in the last column of Tab. 4. However, that only applies to the time investments which leaves those in the financial categories significantly underinvested.

4. Conclusion

The survey data and results are based on an unrepresentative sample so therefore need to be revisited with further research. Nevertheless, the general framework was set to be inductivist; hence some valuable insights still came up. The methodological process, which is both unique and built on previous research, proved to be further replicable as there are significant differences among various impact intensities and cyber and physical matrices. While examining the distribution of frequencies an exciting phenomenon we call “the breaking intervals” emerged. They represent a threshold beyond which individuals are, on average, willing to invest time rather than money or *vice versa*. Relationships among collected variables are what remains of secondary goals to be addressed here. We found that men are slightly more knowledgeable or cautious when it comes to cybersecurity. At the same time we could not find any statistically significant effect of gender, age, education, income or knowledge of cybersecurity basics on the difference between willingness to invest in measures against cyber threats and their physical counterparts.

Table 3. Respondents‘ Willingness to Protect Themselves Against the Impact Intensities.



	CT1	CT2	CT3	CT4	CT5	CF1	CF2	CF3	CF4	CF5	PT1	PT2	PT3	PT4	PT5	PF1	PF2	PF3	PF4	PF5
LH1	18	79	119	29	2	68	112	64	2	1	28	76	84	36	23	35	127	59	17	9
LH2	11	29	110	79	18	33	75	114	23	0	5	31	96	73	45	12	68	108	42	17
LH3	8	18	53	59	109	22	49	71	62	43	5	18	54	63	107	8	39	73	58	69
E1	13	45	133	47	9	35	114	87	10	1	24	92	81	32	18	38	114	54	30	11
E2	7	25	110	72	33	23	71	123	21	9	10	45	89	74	29	25	74	72	52	24
E3	7	15	75	70	80	18	45	105	49	30	10	26	68	73	70	17	46	60	52	72
D1	49	76	100	20	2	89	86	64	4	4	24	157	48	12	6	70	78	71	16	12
D2	9	33	121	68	16	29	91	106	12	9	7	96	96	33	15	25	78	86	40	18
D3	4	16	76	88	63	17	52	106	46	29	6	55	90	51	45	19	38	83	41	66

In terms of the cyber-physical comparison, besides frequencies in Table. 3. with self-explicable results, the focus was on how individuals are willing to invest time or money in measures against cyber and physical threats that are represented by their impact, so are hence comparable. We were also interested in the change between those two conditions (cyber and physical), be it negative, positive or be it a tie. This change of willingness indicates to which threat impact respondents attach more weight. Results were also put in the context of threat analysis undertaken by Jan Kleiner [24].

In the life and health category, on average respondents tended to invest more in physical countermeasures than in cyber ones. That is in accordance with real-world risks. The data category was set to counter-balance the more physical-dominant life and health category, so we had expected an apparent inclination to invest more in cyber countermeasures (e.g. reading cybersecurity tips or books, creating a firm password policy, buying an anti-virus licence) against threats like data and identity theft that are currently on the rise [27]. The trend did not follow our expectations. Individuals were willing to invest more time against cyber than physical threats. However, the financial investments were much higher on the side of physical threats like theft of an ID card or sensitive document. Despite the massive amount of ransomware and other financial attacks all over the world, there are physical, and financial threats such as theft, fraud or embezzlement prevalent in the Czech Republic [26], but the respondents were willing to invest more in countering cyber threats, even though they occur less frequently.

Table 4. Wilcoxon signed-rank test ranks.

Pairs of variables (intensity of impact; type of investment)	Rank (negative= PHYS<CYB; positive=PHYS>CYB)	Life and health impact (number of ranks)	Economic impact (number of ranks)	Data impact (number of rank
PHYS vs. CYB (low; time)	Negative		106	94
	Positive	Not significant	49	58
	Ties		92	95
	Significance	0.071	0.000	0.018
PHYS vs. CYB (medium; time)	Negative	51	90	115
	Positive	87	56	40
	Ties	109	101	92
	Significance	0.000	0.039	0.000
PHYS vs. CYB (high; time)	Negative		82	119
	Positive	Not significant	47	37
	Ties		118	91
	Significance	0.696	0.022	0.000
PHYS vs. CYB (low; financial)	Negative	44	64	35
	Positive	99	71	86
	Ties	104	112	126
	Significance	0.000	0.048	0.000
PHYS vs. CYB (medium; financial)	Negative	38	56	41
	Positive	99	81	82
	Ties	110	110	124
	Significance	0.000	0.001	0.000
PHYS vs. CYB (high; financial)	Negative	60	47	41
	Positive	99	96	96
	Ties	88	104	110
	Significance	0.000	0.000	0.000

Such conclusions are valuable on their own as they can serve as a background for states to more efficiently protect their citizens in cyberspace. Moreover, if we combine the conclusions together, we get a picture, although an incomplete one, of the pungent issue. Individuals from our sample, which is made mostly of higher-educated individuals, are not sufficiently equipped to assess cyber threats on their own. State involvement is thus desirable. It becomes necessary when the logic "a state is only as secure as its citizens" is applied here. Whether there is a call for this is another question requiring further research.

5. Discussion

Our study can serve as a source of valuable information for a state upon which concrete measures can be built. Concretely, encouragements or incentives can be implemented where the willingness is low and savings where it is high. Other efforts like original research, or replication of our

study (with the mitigation its limits) should be made to understand better the recipients of today's and future state's cybersecurity solutions to ensure effectiveness. We see three priorities that arose while conducting our study: a deeper investigation into the "breaking intervals", a deeper investigation into the number of ties, and finding the statistically significant grouping variables which explain the shift between the cyber and physical values of investment.

Despite the significant differences among variables mentioned in the conclusion, the number of ties in the signed-rank test, i.e. people who chose the same investment for the cyber and physical reality, took on values around 100 in all cases. This indicates the lack of contemplation, general knowledge about cyber threats and their risks, or the shortage of physical vs cyber recognition. It could also be caused by an effort to undertake the survey as quickly as possible while 'satisficing', a term used by Schaeffer & Presser [11], the researchers. On the other hand, we do not think that chaotic and meaningless answers can explain the high number of ties as they are a sign of consistency. Chaotic answers would vary much more. It must also be emphasized that "the breaking intervals" are probably closely tied with the sample, or rather its average income and socioeconomic status, how closely, we simply do not know, and it makes "the breaking intervals" another interesting and valuable topic worthy of academic pursuit.

It is also worth viewing our results (with all their limitations in mind) in light of the Kävrestad et al. [7] paper, which concluded that the threat itself might not be the predictor for users' willingness to pay for cybersecurity training as those two variables did not correlate. Our results might suggest the possible explanation of impacts being one of the variables of interest for such studies.

Finally, there are also implications for the practical conduct of a security policy. It has roots in nudge theory as streamlined by Thaler & Sunstein [28] and which concerns, among other things, how to best alter the governance and administrative process so they have the desired effect on citizens. Our results and those of Käverstand's et al. [7] suggest that in communication with citizens, a government should emphasize the impact of cyber threats, not the cyber threats themselves. A possible and established reason for this could be the longstanding lack of cybersecurity knowledge possessed by the average person [29].

Funding

This article was written at Masaryk University with the support of a Specific University Research Grant provided by the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] *KnowBe4, Security Awareness Training* [Online]. Available: https://www.knowbe4.com/en/products/enterprise-security-awareness-training/. [Accessed: Mar. 30, 2020].

[2] Z. Yan, T. Robertson, R. Yan, S. YonPark, S. Bordoff et al., "Finding the weakest links in the

weakest link: How well do undergraduate students make cybersecurity judgment?," *Computers in Human Behavior*, vol. 84, pp. 375–382, 2018, doi: 10.1016/j.chb.2018.02.019.

[3] BBC. (2018, Oct. 5). *Weak passwords banned in California from 2020* [Online]. Available: https://www.bbc.com/news/technology-45757528. [Accessed: Nov. 3, 2022].

[4] E. Dreyfuss. (2018, Oct. 6). *A Good Password Law, Hardware Hacks, and More Security News This Week* [Online]. Available: https://www.wired.com/story/security-news-this-week-good-news-california-bans-bad-default-passwords/. [Accessed: Nov. 3, 2022].

[5] Williams D. (2019, Feb. 18). *Israeli cyber-hotline offers help for the hacked, Reuters* [Online]. Available: https://www.reuters.com/article/us-cyber-israel-hotline-idUSKCN1Q70K1. [Accessed: Nov. 3, 2022].

[6] B. Willemsen and M. Cadee, "Extending the airport boundary: Connecting physical security and cybersecurity," *Journal of Airport Management*, vol. 12, no. 3, pp. 236–247, 2018.

[7] J. Kävrestad, M. Gellerstedt, M. Nohlberg, and J. Rambusch, "Survey of Users' Willingness to Adopt and Pay for Cybersecurity Training," in *Human Aspects of Information Security and Assurance: 16th IFIP WG 11.12 International Symposium, HAISA 2022*, Greece, N. Clarke, S. Furnell, Eds. Cham: Springer, 2022, pp. 14–23. doi: 10.1007/978-3-031-12172-2_2.

[8] S. M Furman, M. F. Theofanos, Y. Choong, B. Stanton, "Basing Cybersecurity Training on User Perceptions," *IEEE Computer and Reliability Societies*, vol. 10, no. 2, pp. 40–49, 2012, doi: 10.1109/MSP.2011.180.

[9] S. Bauman, M. L. Newman, "Testing assumptions about cyberbullying: Perceived distress associated with acts of conventional and cyber bullying," *Psychology of Violence*, vol. 3, no. 1, pp. 27–38, 2013, doi: 1037/a0029867.

[10] M. Fagan, M. M. H. Khan, "To Follow or Not to Follow: A Study of User Motivations around Cybersecurity Advice," *IEEE Computer Society*, vol. 22, no. 5, pp. 25–34, 2018, doi: 10.1109/MIC.2017.3301619.

[11] N. C. Schaeffer, S. Presser, "The Science of Asking Questions," *Annual Review of Sociology*, vol. 29, pp. 65–88, 2003, doi: 10.1146/annurev.soc.29.110702.110112.

[12] J. Bourke, A. Kirby, and J. Doran, *Survey and Questionnaire Design*. Ireland: Oak Tree Press, 2016.

[13] F. J. Fowler and T. W. Mangione, *Standardized Survey Interviewing: Minimising Interviewer-Related Error*. Newbury Park, CA: Sage Publications, 1990.

[14] J. Kleiner. (2016). "Analýza kybernetických hrozeb eGovernmentu a jejich rizik pro ČR" [Online]. Available: https://is.muni.cz/th/inn08/KLEINER_Analyza_horzeb_eGovernmentu_a_jejich_rizik_pro_CR.pdf. [Accessed: Nov. 7, 2022].

[15] P. Zeman, *Česká bezpečnostní terminologie: výklad základních pojmů*. Brno: Masaryk University, 2002.

[16] Techopedia. (2022, Apr. 25). *Definition – What Does Cyberthreat Mean?* [Online]. Available: https://www.techopedia.com/definition/25263/cyberthreat. [Accessed: Nov. 7, 2022].

[17] *Merriam-Webster, Definition of physical* [Online]. Available: https://www.merriam-webster.com/dictionary/physical. [Accessed: Nov. 8, 2022].

[18] E. P. Cox, "The optimal number of response alternatives for a scale: a review," *Journal of Marketing Research*, vol. 17, no. 4, pp. 407–422, 1980, doi: 10.1177/002224378001700401.

[19] Český statistický úřad. (2019). *Průměrné mzdy – 1. čtvrtletí 2019"* [Online]. Available: https://www.czso.cz/csu/czso/cri/prumerne-mzdy-1-ctvrtleti-2019. [Accessed: Nov. 7, 2022].

[20] Czech Republic. (2009). *Zákon č. 40/2009 Sb. Zákon trestní zákoník* 2009 [Online]. https://www.zakonyprolidi.cz/cs/2009-40/zneni-20220901. [Accessed: Nov. 7, 2022].

[21] W. Chai. (2022). *Definition: confidentiality, integrity, and availability (CIA triad)* [Online]. Available: https://whatis.techtarget.com/definition/Confidentiality-integrity-and-availability-CIA. [Accessed: Nov. 7, 2022].

[22] Google, *Tipy, jak zajistit větší online bezpečnost* [Online]. Available: https://safety.google/intl/cs/security/security-tips/. [Accessed: Nov. 7, 2022].

[23] UC Berkeley, *Top 10 Secure Computing Tips* [Online]. Available: https://security.berkeley.edu/resources/best-practices-how-to-articles/top-10-secure-computing-tips. [Accessed: Nov. 7, 2022].

[24] J. Kleiner. (2020). *Kybernetická bezpečnost koncových uživatelů v ČR a jejich ochota se zabezpečit* [Online]. Available: https://is.muni.cz/th/xsa0b/Kleiner_DP_Kyberbezpecnost_koncovych_uzivatelu_Archive.pdf. [Accessed: Nov. 7, 2022].

[25] Visions of Humanity. (2020). *Global Peace Index 2020* [Online]. Available: https://www.visionofhumanity.org/maps/#//. [Accessed: Nov. 7, 2022].

[26] D. Schimmer. (2019). *Statistika kyberkriminality* [Online]. Available: https://www.policie.cz/clanek/statistika-kyberkriminality.aspx. [Accessed: Nov. 8, 2022].

[27] T. A. Tunggal. (2020, June 1). *The 36 Biggest Data Breaches* [Online]. Available: https://www.upguard.com/blog/biggest-data-breaches. [Accessed: Nov. 8, 2022].

[28] R. H. Thaler and C. R. Sunstein, N*udge: Improving decisions about health, wealth, and happiness.* New Heaven: Yale University Press, 2008.

[29] K. Olmstead, A. Smith, (2017, Mar. 22). *What the Public Knows About Cybersecurity* [Online]. Available: https://www.pewresearch.org/internet/2017/03/22/what-the-public-knows-about-cybersecurity/. [Accessed: Nov. 8, 2022].

Privacy on the Internet: An Empirical Study of Poles' Attitudes

Daniel Mider | **Faculty of Political Science and International Studies, University of Warsaw, Poland**, ORCID: 0000-0003-2223-5997

Abstract

The value system of Poles in terms of the phenomenon of privacy on the Internet was analysed. The following aspects were taken into account: privacy on the Internet as a moral value, privacy on the Internet as a subject of legal regulations (current or future) and actual actions taken by users to protect privacy. The differentiation of Polish society in terms of the three above-mentioned areas was also examined. Results were obtained on the basis of a quantitative empirical study conducted on a representative sample (N=1001) of adult Poles. The method of computer assisted telephone interviews (CATI) was used. Descriptive statistics and selected inductive statistics were used in the analyses. Intra-group differentiation was investigated using a method called two-step cluster analysis. Poles have low technical competences in the field of Internet privacy protection. This value is appreciated; however, it rarely translates into active protection of one's own identity and information. A strong polarization of Poles' attitudes towards the requirement to disclose their identity on the Internet was identified, as well as ensuring access to any user information by law enforcement agencies. Poles are willing to accept legal regulations preventing their profiling. We note a moderately strong negative attitude towards state institutions as a factor limiting privacy on the Internet and a significantly lower (but still negative) attitude towards Internet service providers. Poles differ in terms of attitudes towards privacy on the Internet (IT competences, age, education, gender, socioeconomic status and size of the place of residence).

Corresponding author:
Daniel Mider, Faculty of Political Science and International Studies, University of Warsaw, Poland, ORCID: 0000-0003-2223-5997; d.mider@uw.edu.pl

Keywords

online behavior, online freedom, privacy paradox, privacy perceptions

Cite this article as: D. Mider, "Privacy on the Internet: An Empirical Study of Poles' Attitudes," ACIG, vol. 1, no. 1, pp. 253–271, 2022, DOI: 10.5604/01.3001.0016.1459

1. Introduction

IN this article, we analysed the value system of Poles in terms of the phenomenon of privacy on the Internet. The following aspects were considered: privacy on the Internet as a moral value, privacy on the Internet as a subject of legal regulations (current or future), and actual actions taken by users to protect privacy. The differentiation of Polish society in terms of the three above-mentioned areas was also examined, creating its segmentation.

We are constantly witnessing data privacy violations of ordinary Internet users. These threats are multi-vector. Cybercriminals are an obvious source of threats, although awareness of the scale of these threats is low among ordinary users. As an exemplification, it can be mentioned that the leaked databases, available to anyone interested who is willing to pay, often have over 20 billion records (i.e., user login and password pairs, directly or in an encrypted form) [1]. Another vector of threats is Big Tech. Global social media operators collect data excessively and this data is used – directly or indirectly – to manipulate social groups, as demonstrated by whistleblowers Christopher Wylie and Brittany Kaiser, former employees of Cambridge Analytica. The non-obvious entity violating user data is national states that implement surveillance programs which collect data in an oppressive manner. The best-known example is the US National Security Agency's surveillance system, code-named PRISM, which was disclosed by Edward Snowden [2]. In the study entitled "The Future of Privacy" conducted by the Pew Research Center in 2014, over 2,500 experts expressed a pessimistic view that privacy on the Internet would disappear by 2025 [3].

In the context of the above-mentioned facts, the research problem was defined by posing the following research questions:

- What is the level of Poles' awareness of violations of their privacy on the Internet?
- What are the predictors of differences in attitudes towards online privacy: sociographic, psychographic, or behavioural characteristics?
- What are their attitudes towards possible legal regulations regarding restrictions or extensions of privacy protection?
- Do they practically protect their privacy on the Internet and to what extent?
- Do they want their privacy to be protected by themselves, or do they prefer the obligation to protect to be transferred to another entity?

The above-mentioned problems are listed in the area of interest of numerous researchers. These considerations omit numerous legal sociological and IT references, focusing primarily on the current empirical findings in the field of attitudes towards the phenomenon of privacy. An important field of interest is the awareness of privacy risks and ways of understanding this concept in the context of functioning on the Internet.

Above all, the speed of changing technology is emphasized, and as a result, it is difficult to precisely define the concept of privacy. According to some researchers, this concept is almost impossible to define [4]. Research on privacy shows a low level of awareness of this phenomenon by societies. On the other hand, regardless of the superficial understanding of privacy issues, it is widely appreciated. The respondents blame the violation of privacy, while understanding the value of sharing information. However, they do not want the sharing of information against their will and intentions [5]. To date, the most comprehensive and cited privacy survey was conducted by an American public opinion poll Pew Research Center [6]. A valuable and interesting empirical study was also the EMC Privacy Index, which over time transformed into the Dell Technologies Global Data Protection Index [7]. The latter, however, has evolved towards cybersecurity issues (e.g., threats such as phishing or ransomware). Comprehensive research of privacy on the Internet has not been carried out in Poland so far. This fact became the basis for undertaking the examination of this issue. Research on the broadly understood privacy appeared in Poland sporadically and referred to the degree of measurement to which various values are held. Two such studies have recently been carried out: relating to the extent to which Poles value and are willing to protect their data [8], and a study by the Center for Studies on Democracy of the SWPS University on the value of privacy and freedom on the Internet [9]. However, the raw data was not released – the research is available through press articles and in the form of short reports.

Academic studies of socio-demographics which correlates attitudes towards privacy are commonly studied. First of all, researchers focus on the gender category [10–12]. Conclusions were also formulated in relation to the age of the respondents, indicating groups particularly sensitive to threats of privacy: adolescents [13, 14] and seniors [15].

Researchers' efforts also focus on the phenomenon of the correlation between awareness of threats to privacy and the lack of importance; or even neglect of its protection by users. This phenomenon has been referred to as the privacy paradox [16–18]. This particularly applies to the use of smartphones [19, 20]. It has been shown that online behaviour on the Internet often resembles the behaviour of users with low IT competences [21].

The interest in the issue of privacy in the academic and political environment, especially in the perspective of the above-mentioned threats, has been attracting attention since the 19th century. The cornerstone of the modern debate on privacy is an article by American lawyers Samuel D. Warren and Louis D. Brandeis [22]. The right to privacy is placed among the first-generation human rights and is subject to universal regulation at the level of international and national legislation (Universal Declaration of Human Rights, International Covenant on Civil and Political Rights, Convention for the Protection of Human Rights and Fundamental Freedoms). It is also worth adding that one of the earliest topics related to the ethics of new technologies that aroused public interest was privacy. In the mid-1960s, the US government created large databases containing information about citizens (these were census data, tax records, military service records, and social records). It was then that the first public debate

on limiting the government's appetite for information about citizens was initiated. Another discussion began in the 1980s as a result of the development of information technology. Continuing that a social movement was formed, the doctrinal basis of which was the belief that the right to privacy was defective by state institutions. It was argued that states have violated the requirement to care for this value and in this respect pose the greatest threat to the citizen. Moving forward the term Privacy Enhancing Technologies (PET) was created, meaning such technical solutions that provide users with complete privacy and exclusive control over the data they create and send. The perceived threats have now led to the development of the concept of individual digital self-determination. This concept assumes that the interactions in cyberspace, especially by large entities: should be transparent, data should not be excessively collected, and the user should not be manipulated on the basis of algorithms and information unintentionally left in cyberspace (metadata). In addition, everyone should independently manage information about themselves and decide who and under what conditions has the right to access it. Currently, it is postulated that these freedoms should be introduced into the Charter of Fundamental Rights of the European Union.

For the purposes of this text: the concept of privacy and the concept of anonymity, and those closely related to it, have been defined (bearing in mind that these terms are considered widely in academic publications) [23]. Privacy is understood as a situation where everyone knows our identity, but no one knows what we are doing, and therefore what data we exchange. On the other hand, anonymity is a situation where no one knows who we are (so they do not know, for example, our name and surname, or any other information about us that can reveal who we are); however, everyone can see our actions. It is emphasized that the two concepts connect with each other, because in numerous online activities the possibility of maintaining privacy without anonymity is difficult or even impossible to implement [24].

2. Methods

The research questions were answered during a quantitative empirical study conducted by the Association of Political Science Graduates affiliated with the Faculty of Political Science and International Studies at the University of Warsaw. The research was financed by the Justice Fund and administered by the Minister of Justice. The measurement entitled *Premises of a sense of security. Privacy, anonymity, freedom, and security* were conducted from December 1st to 23rd, 2021 using the computer-assisted telephone interview method as a representative (gender, age, size of the place of residence, and education) sample of N=1001 adult (18+) Poles. The measurement, fulfilling the requirements of the so-called statistical representativeness, can be generalized from the sample to the population of adult Poles. The maximum standard error of the estimation was ±3.1%.

The measurement was carried out using the Computer Aided Telephone Interviews (CATI) technique. This technique has been considered better than classic standardized face-to-face interviews (Paper and Pencil Interviews, PAPI) and online surveys (Computer Assisted Web Interviews,

CAWI). In relation to the above-mentioned techniques, CATI has numerous methodological, psychological, interactional, organizational, and technical advantages which made it a useful tool for this project. As regards to the methodology, it involves a higher accessibility of respondents and a higher availability of the sampling frame (as compared to other methods) corresponding to general population. In terms of the psychological and interactional aspect, the comfort of interaction between the interviewer and the respondent is significantly greater. Communication over the phone distinctly increases the sense of anonymity for the respondent, which translates into respondents' feeling more at ease about expressing their views on difficult or sensitive issues. In terms of the technical and organizational aspect, an important feature of telephone interviews is the high level of control over the research process. This concerns both the human factor (interviewers and respondents), as well as the collected data. Also, the use of computer programmes and telephone contact significantly decrease the financial and organizational costs necessary to carry out the research. As a result, we achieve a higher response rate as compared to other research methods, higher quality of data which are precise and accurate, and a low level of errors, as well as reliability and accuracy.

A *sine qua non* condition of generalization of the results from the sample to the studied population is sampling (i.e., simple random sampling) according to the standards, and its sufficient size. Sampling will be made with the use of the method ensuring the randomness of sampling, developed as part of US state methodology of quantitative research (and widely adopted in research practice). The procedure was developed by Warren Mitofsky and Joseph Wakesberg. It is referred to in research practice as Random Digit Dialling (RDD) [25], and among researchers using computer-assisted telephone interviews it is considered an optimal and classic method [26].

The scope of the research procedure carried out included the assessment of attitudes towards the phenomenon of privacy in the following three aspects: moral, legal, and behavioural.

The moral aspect of attitudes towards privacy on the Internet. It is understood as the ethical limits of protecting and concealing one's identity on the Internet, set from the perspective of ordinary users, i.e., the weakest entities. In order to test this aspect, the following three questionnaires were selected from the research tool:

- M1. Edward Snowden is a former employee of the American intelligence (CIA) who in 2013 disclosed classified information regarding numerous global surveillance programs of the USA, undertaken by the country in cooperation with companies and some European countries. How do you evaluate such a phenomenon of mass surveillance by states on the Internet?

It was a closed, one-answer question. The respondent could indicate one of the following five answers: I strongly support, I rather support, I rather consider it unacceptable, I consider it completely unacceptable, and I have no opinion on this subject (the last item was not read by the interviewer). Such a scale was chosen due to the possibility of treating the variable as ordinal and, as a result, calculating such statistics as the mean

or standard deviation. The next two questions used as indicators of the moral aspect were as follows:

- M2. Profiling and tracking us online by service providers is simply the price we should pay for our convenience.
- M3. Anyone who wants to be anonymous on the Internet is either a cybercriminal or has bad intentions.

Questions M2 and M3 were closed-ended single-answer questions. The respondent could choose from the following responses: strongly disagree, rather disagree, rather agree, strongly agree.

Legal aspect of attitudes towards privacy on the Internet. Views on support for legal compulsion to disclose one's identity in all online interactions, the possibility of deanonymising and decrypting any private information online by law enforcement, and the legal regulation of user profiling by content providers on the Internet were examined. The following questions were asked:

- L1. Law enforcement authorities should, in important situations, be able to access any of our information on telephones, computers or the Internet, no matter how secure it is.
- L2. Profiling and tracking by service providers on the internet should be prohibited by law.
- L3. Everyone on the Internet should use in all interactions their first and last name, and this should be prescribed by law.
- L4a. The current legal provisions ensure adequate protection of people's privacy in their online activities.
- L4b. The current provisions on the protection of personal data, and therefore, above all, the GDPR, ensure adequate protection of people's privacy in their online activities.

Questions L4a and L4b were a special case. The sample was divided into two parts (the savings were dictated by the research costs), respectively L4a obtained n=500, and L4b n=501 answers. As part of the moral and legal aspect, attitudes towards the institutions of the first sector; i.e., the state (questions M1 and L1), and the second sector; i.e., enterprises owning social media (questions M2 and L2), were also taken into account. In the Results part, the statements were reformulated so as to be able to present the respondents' answers in a homogenous manner (direction of variables).

Behavioural aspect of attitudes towards privacy on the Internet. In this aspect, it was examined which Internet protection measures are actually taken by the respondents. It strictly depends on IT competences. The following question was asked:

Please consider which of the following ways to protect information on the Internet are undertaken by you...

Protection of identity and sensitive information on the Internet is not easy, so the respondent could choose from the following list of indicators: B1. I delete documents/files that should not fall into the wrong hands. B2. I am using the browser in incognito mode. B3. I use aliases so that I cannot be traced back to my real name. B4. I delete cookies/delete browser history. B5. I use a temporary username or email. B6. I encrypt files/documents. B7.

I use Virtual Private Networks (VPNs). B8. I exercised the so-called "right to be forgotten". B9. I use advanced anonymisation tools (e.g., The Onion Router – Tor, Invisible Internet Project – I2P, Linux TAILS, Linux Whonix, OTR communication encryption).

These were dichotomous, one-answer questions, for each of them the respondent could answer "yes" or "no". The security measures indicated in the questions measured the level of technical and IT advancement of the respondent, including also verifying activities that could be considered wrong, not contributing to the strengthening of privacy. Items B1 and B2 are among the counter-effective actions (myths, mistakes). B3, B4 are among the elementary, basic, and partially effective actions. Activities that can be described as intermediate are: B5, B6, B7. By contrast, advanced privacy measures include questions B8 and B9.

In order to indicate social, demographic, and psychographic predictors of individual attitudes; inductive statistics were used. Chi-square is used to determine whether or not there is a relationship between the variables. In order to find out how strong a relationship is (expressed in the range from zero to one), the Harald Kramer's statistics (V) and the Karl Pearson's contingency coefficient (C) were used.

Cluster Analysis. In the next step of the research procedure, it was checked what types of attitudes towards privacy on the Internet can be distinguished in Polish society. For this purpose, segmentation was made in order to discover the smaller structures of Polish society in terms of attitudes towards privacy.

The premise for the use of this statistical method is the necessity to reduce data resulting from the multi-faceted nature of privacy. Cluster analysis is a group of diverse statistical techniques used to classify cases into groups that are relatively homogeneous within themselves and heterogeneous among themselves. These groups are called clusters. Cluster analysis is the so-called unsupervised learning method – "without a teacher". This method is opposed to discriminant analysis (supervised classification). In unsupervised classification, the group structure does not have to be known *a priori*. This makes cluster analysis attractive as an exploratory tool. Cluster analysis detects structures in data without explaining why these structures exist. It is a method concurrent with human intuitive, everyday reasoning; which consists on grouping objects on the basis of similarity.

The method was invented in anthropology by Harold E. Driver and Alfred L. Kroeber in 1932 [27]; although the need for such a data mining technique was previously expressed by a Polish scientist: a supporter of statistical studies in anthropology, Jan Czekanowski [28]. It was popularized in science by Raymond B. Cattel using for the classification of personality traits [29]. The career of this method began in the 1960s and 1970s [30]. It has stimulated worldwide research into clustering methods and has initiated numerous publications on the subject; furthermore, it is widely used in various scientific disciplines [31].

Specifically, Two-Step Cluster analysis, was used. This analytical technique has particularly useful features: the ability to construct a model using both interval and nominal variables, and it allows the analysis of databases with large numbers of units of analysis. The input data finally selected for the segmentation performed were the 8 previously

mentioned attitudes and behaviours of users: M1, M2, M3, L1, L2, L3 and B9. Additionally, the behavioural variable has been enabled: I use pseudonyms on the Internet so that they cannot be associated with my real name (possible answers "yes" or "no").

3. Results

The following analytical aspects are presented below: moral attitudes towards privacy on the Internet, attitudes towards legal regulations of this phenomenon (including the assessment of first and second sector institutions), and behavioural aspects of privacy protection by ordinary Internet users. The analytical part ends with segmentation: groups of Poles with different attitudes towards the phenomenon of privacy on the Internet have been identified.

3.1. Moral attitudes of Poles towards privacy on the Internet

Poles represent various attitudes towards the moral aspects of privacy on the Internet. More than three-quarters of them (82.1%) have a negative opinion of the mass surveillance system codenamed PRISM. On the other hand, only slightly more than a quarter of them (28.2%) consider it acceptable that companies providing services on the Internet collect excessive amounts of data from users. Opinions about the value of anonymity are divided. It should be emphasized that almost two-thirds of Poles (62.1%) recognize that anonymity on the Internet is a positive value and does not have to serve unethical or even criminal activities.

Table 1 Moral subaspects of attitudes towards privacy on the Internet

	Response rate (in %)		
Statements	**Negative responses (negative attitudes towards privacy in cyberspace**	**Undecided**	**Affirmative responses (positive attitudes towards privacy in cyberspace)**
M1. Assessment of the phenomenon of mass surveillance by state entities (E.J. Snowden's case)	10.3	7.6	82.1
M2. User convenience is not enough to pay for online profiling and tracking by service providers	50.8	11.0	28.2
M3. Wanting to remain anonymous does not mean being a criminal or having malicious intent	29.7	8.2	62.1

The fact that more than half of service providers consider the activities of service providers on the Internet regarding data collection as ethical will undoubtedly hinder both the introduction of legal solutions limiting these providers that regulate this market, as well as the public discussions on this subject. The knowledge of Internet users about the unethical activities of corporations in the field of obtaining excessive amounts of data and the unethical methods of their use is still small.

Those located in the political center, center-left, or center-right express particularly strong opposition to government surveillance. In addition, people who do not have specific political views or who define themselves as off-scale left-right, and therefore probably people with mixed or libertarian views, are negative about state surveillance [χ^2 (24, N=1001) = 49.53; p≤0.01; V=0.131; C=0.22]. Predictors of positive attitudes towards privacy on the Internet are living in a medium or large city; i.e., over 50,000 inhabitants, but this value is on the border of statistical significance [χ^2 (21, N=1001) = 29.75; p≤0,1; V=0.12; C=0.20]. The subjective sense of economic status is a weak but statistically significant correlate of positive attitudes towards privacy on the Internet. It is non-linear; i.e., the opposition to government surveillance is higher among those who indicate average or moderately high income [χ^2 (15, N=1001) = 30.49; p≤0.01; V=0.11; C=0.18]. The lowest percentage of opposition to government surveillance was recorded among those who say they have insufficient money for even the cheapest food. Opposition to government surveillance is most strongly correlated with political attitudes identified on the basis of self-identification. The premise of negative attitudes towards collecting excessive amount of data by corporations is age. We can see a positive correlation here, so, the younger the respondent, the more likely he is to accept corporate data collection [χ^2 (1, N=1001) = 10.51; p≤0.001; V=0.13; C=0.21].

The higher the age [χ^2 (1, N=1001) = 133.19; p≤0.001; V=0.24; C=0.38], the smaller the size of the place of residence [χ^2 (1, N=1001) = 5.63; p≤0.001; V=0.14; C=0.24], the less consent to moral justification of anonymity on the Internet. Moreover, we find the most positive attitudes towards anonymity on the Internet in people who describe themselves as right-wing [χ^2 (24, N=1001) = 51.64; p≤0.001; V=0.13; C=0.22].

3.2. Poles' attitudes towards legal regulations concerning privacy on the Internet

The results presented in Tab. 2. reveal the polarization of views regarding the regulation of privacy on the Internet; we observe a low level of libertarian attitudes. A moderate consensus is possible regarding the limitation of data collection by economic entities managing social media. In this case, the majority (60.0%) of the respondents would be willing to accept the legal limitation of this phenomenon.

Table 2. Legal subaspects of attitudes towards privacy on the Internet – changes to law

	Response rate (in %)		
Statements	**Negative responses (negative attitudes towards privacy in cyberspace)**	**Undecided**	**Affirmative responses (positive attitudes towards privacy in cyberspace)**
L1. Law enforcement authorities should not be able to access each of our information, even in important situations	46.8	9.4	46.6
L2. We must be legally prohibited from being tracked and profiled by service providers on the Internet	30.5	9.4	60.0
L3. The law should not require you to appear on the Internet only under your first and last name	44.0	6.7	49.3

A statistically significant and moderately strong correlation was observed between supporting the idea of legal regulations and the age of the respondents [L1. χ^2 (1, N=1001) = 9.48; p≤0.001; V=0.12; C=0.21; L2. χ^2 (1, N=1001) = 4.21; p≤0.05; V=0.16; C=0.30; L3. [χ^2 (1, N=1001) = 34.13; p≤0.001; V=0.17; C=0.32]. The phenomenon is especially intensified in the group over 55 years of age. Gender is a weak correlate of support for legal regulations, but it was treated as a phenomenon accompanying advanced age. The right-wing (62.7%) and people with unspecified political views (54.8%) would be particularly eager to grant powers to dispatching services [χ^2 (24, N=1001) = 77.61; p≤0.001; V=0.17; C=0.28]. The desire to limit the possibility of collecting data by corporations is particularly visible among people who define themselves as left-wing (64.3%), and also among the right-wing (40.5%) [χ^2 (32, N=1001) = 66.81; p≤0,001; V=0.13; C=0.25]. The remaining socio-demographic variables turned out to be statistically insignificant.

The General Data Protection Regulation (GDPR) is a legal act in force in the EU Member States. In Poland, it entered into force on May 25, 2018. Despite the fact that it significantly protects privacy, it was initially viewed negatively, in particular by entrepreneurs. The sample was randomly divided into two parts: In the first group, question L4a (general assessment of privacy rights) was asked; in the second group, question L4b (direct reference to the provisions of the GDPR). Such a procedure was aimed at checking whether the current social attitude towards the GDPR has changed and whether the use of the trigger-word; i.e., "GDPR", is a reason for the occurrence of differences in the distribution of responses.

Table 3. Legal subaspects of attitudes towards privacy on the Internet – current law

	Response rate (in %)		
Statements	**The protection afforded by law is excessive**	**Undecided**	**The protection afforded by law is too weak**
L4a. The current legal provisions ensure adequate protection of people's privacy in their online activities	12.4	21.1	66.5
L4b. The current legal provisions on the protection of personal data, and therefore, above all, the GDPR, ensure adequate protection of people's privacy in their online activities	16.3	13.9	69.8

It was shown that the differences between the distribution of answers to both questions are statistically significant ($p \leq 0.05$), but the differences in the response rates are not large. In the case of the answer that the legal protection of privacy is too extensive, a difference of only slightly less than 4.0% was noted.

3.3. Behavioural aspects of privacy protectionon the Internet

As shown in Tab. 4., the measures most often taken by Poles to protect privacy are not very effective or completely ineffective. Most of all, deleting cookies (which is done by more than half of the respondents (54.8%)), deleting documents or files so that they do not fall into the wrong hands (49.1%), and using web browsers in *incognito* mode (31.9%). Poles make little use of effective legal solutions, such as the right to be forgotten, or of effective IT solutions such as anonymisation tools, and thus ensuring almost complete security.

The analysis of sociodemographic variables showed that only one factor, that is age, correlates with the competences in the field of privacy protection on the Internet. The probability of taking counter-effective or only partially effective measures increases with age. This applies to deleting documents [χ^2 (5, N=1001) = 81.35; p≤0,001; V=0.29; C=0.28] and using the browser in incognito mode [χ^2 (5, N=1001) = 50.86; p≤0,001; V=0.22; C=0.21]. Education is an important predictor of undertaking most of the activities protecting privacy. The higher the education, the more actions are taken such as: document encryption [χ^2 (8, N=1001) = 53.14; p≤0,001; V=0.23; C=0.23], using a VPN [χ^2 (8, N=1001) = 37.85; p≤0,001; V=0.19; C=0.19], using web browser in incognito mode [χ^2 (8, N=1001) = 39.11; p≤0,001; V=0.20; C=0.19], and deleting cookie files [χ^2 (8, N=1001) = 57.56; p≤0,001; V=0.24; C=0.23]. Among other observations, it is also worth pointing out the gender variable – men have a statistically significantly higher tendency to use the browser in incognito mode [χ^2 (1, N=1001) = 12.40; p≤0,001; V=0.11; C=0.11] and using a VPN [χ^2 (1, N=1001) = 24.43; p≤0,001; V=0.16; C=0.16].

Table 4. Behavioural subaspects of attitudes towards privacy on the Internet

	Response rate (in %)	
Statements	**Does not perform / does not use**	**Performs / uses**
B1. Deleting documents/files so as not to fall into the wrong hands	50.9	49.1
B2. Using the browser in incognito mode	68.1	31.9
B3. Using aliases so not to be traced back to real name	72.5	27.5
B4. Delete cookies/Delete browser history	45.2	54.8
B5. Using a temporary username or email	78.6	21.4
B6. Encryption of files/documents	79.0	21.0
B7. Using a VPN	82.1	17.9
B8. Using the so-called "right to be forgotten"	99.4	0.6
B9. Using advanced anonymisation tools	90.4	9.6

3.4. Diversification of Poles' attitudes towards online privacy – segmentation

Variables were selected to create segmentation which enabled the most effective division of the surveyed Poles into clusters. The segmentation carried out using the cluster analysis method led to the identification of three types of attitudes towards privacy on the Internet. The obtained segmentation results turned out to be statistically satisfactory as measured by the coherence and distinctness measure (the so-called Silhouette coefficient). This coefficient indicates whether the division was made in such a way that observations are concentrated within the groups and separated between them. It takes values from – 1 (very weak model) to 1 (perfect model). Fig. 1. shows the fit of the model graphically. The obtained result should be considered at least satisfactory, i.e. almost 0.4.

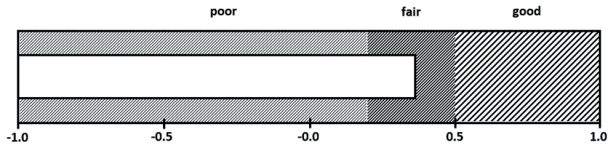


Figure 1. Silhouette measure of consistency and distinctiveness.

The most significant creative task in the procedure of segmentation is to give appropriate names for each of the segments identified. After analysing the characteristics, three descriptive names of the separate groups were created (in order of group size):

- Cluster 1. Skeptical or disappointed praetorians with low or medium IT competences (68.7%; n=668).
- Cluster 2. Moderate cyberlibertarians with medium or low IT competences (21.7%; n=217).
- Cluster 3. Moderate or extreme cyberlibertarians with high IT competences (9.6%; n=96).

The socio-demographic differences shown below are dominant; i.e., the characteristics are presented which make a specific group statistically significantly different from the average value for the population.

Cluster 1. Skeptical or disappointed praetorians with low or medium IT competences. They were called praetorians because they showed a significantly higher than the population average and lower than in other groups predilection towards legal restriction of privacy on the Internet. In particular, they advocate the need to verify with your name on the Internet and the possibility of law enforcement access to any data. Moreover, in the moral aspect, they mostly consider anonymity as being a cybercriminal or having malicious intent. The second premise for calling them praetorians was the fact that this group has the lowest intensity of negative attitudes towards state institutions. This group presents these attitudes as not too intense, and the differences between the other groups are small and in the above-mentioned aspects they do not exceed 15 percentage points. For this reason, the term skeptical or disappointed is used. This group generally equates their internet security with pseudonymisation. Nobody in this group uses advanced anonymisation tools. This group includes mainly women, they are people over 55, mostly inhabitants of rural areas and small towns up to 20,000 inhabitants. These people usually have secondary education, and consider their own material well-being to be moderate. From a psychographic point of view, we see an overrepresentation of practicing believers or non-practicing believers and representatives of the right-wing worldview.

Cluster 2. Moderate cyberlibertarians with medium or low IT competences. In this publication, cyberlibertarians are defined as those who oppose legal solutions to limit privacy on the Internet. The adjective "moderate" was used due to the fact that the differences, although statistically significant, are not so large in nominal terms. In this group, we see in the moral aspect, strong support for anonymity, stronger than in all other

groups, and in the legal aspect, reluctance to impose the necessity to identify Internet users. In this group, we see the greatest consent to violating privacy by corporations. This consent is the greatest in comparison to the other groups and the average for the entire population. IT competences in this group are low, as in the first group, they consider pseudonymisation to be a sufficient way to protect privacy on the Internet. Also, no one from this group uses advanced anonymisation tools. This group is dominated by men, people from 25 to 44 years of age prevail. We see a positive high perception of economic welfare, which is also correlated with the highest actual wages. These people make clear political self-identification. They are of three groups in the following order: left, right, and center. Therefore, there are no undefined and undecided people in terms of worldview, as well as center-left or center-right people.

Cluster 3. Moderate or extreme cyberlibertarians with high IT competences. In this group, we see the highest aversion towards state institutions among other groups, although the aversion towards the second sector institutions in terms of violating the privacy of users on the Internet is similar to that in other groups. IT competences within this group are the highest, each of its representatives uses advanced nonymization tools. This group is dominated by men, they are people between 35 and 44 years of age. We see an overrepresentation of the inhabitants of cities with more than 500,000 inhabitants and more than 100,000 inhabitants, as well as the inhabitants of the Mazowieckie Voivodeship. These are people who generally have higher education, who are in cohabitation or who declare themselves as single. They positively assess their own material well-being. They consider themselves as non-believers and consider themselves off-scale left versus right. It is worth noting that in all three groups, the assessment of the phenomenon of mass surveillance by state entities (E. Snowden case) and moral attitudes towards corporations (and therefore consent to the claim that profiling and tracking on the Internet is a payment for free services) are the same.

4. Discussion

In the first place, it is necessary to answer the last of the research questions that were asked in the introduction. This question concerns whether Poles want to protect their privacy on the Internet on their own, or whether they expect protection from entities such as the state or corporations in this respect. It should be emphasized that the attitudes of Poles towards freedom are ambivalent. According to CBOS, freedom is in one of the last places among Poles' other values [32]. Poles primarily value health, family, and work; while only 3.0% of them spontaneously indicated freedom as an important value. A survey conducted by CBOS a year later reveals where the value of freedom was asked directly, reveals that for nine out of ten Polish citizens it is a key value in their lives [33] (this result can be partially explained by the strong political polarization of Polish citizens and the numerous protests that took place at that time). Therefore, this empirical study also posed the question of freedom, but in a different context. The study asked which is more important: freedom or security. However, more

than a fourth of Poles (28.5%) indicate freedom, and an almost identical group (27.8%) claim that security is more important to them than freedom. Most importantly, one third of the respondents (36.2%) indicated that there is security only when people have freedom. As shown by the segmentation, partial values (such as privacy) and closely related to freedom, although undoubtedly valued, remain detached from reality. Firstly, Poles are not aware of the numerous threats to their privacy. Secondly, the behavioural aspect clearly shows that the respondents are not able to protect themselves against privacy violations on their own. Only one in ten respondents is able to effectively maintain privacy on the Internet using both adequate and advanced (effective in all situations) measures.

The conducted analyses showed that in the Polish society it is possible to distinguish groups that differ significantly in their attitudes towards privacy on the Internet. These differences are not of a fundamental nature, as the behavioural aspect resulting from competences turns out to be the most differentiating. Another important factor in distinguishing the group were different attitudes towards anonymity in the moral aspect and different attitudes towards the legal requirement to present one's real identity in online interactions. Moreover, the respondents shared the attitude towards state institutions as potentially violating their privacy on the Internet. The majority of Poles remain relatively skeptical about legal regulations concerning privacy on the Internet. In contrast, attitudes towards second sector entities that potentially breach privacy were slightly varied and only moderately critical. Research also indicates numerous differences in terms of gender and attitudes towards privacy on the Internet. They were taken into account in the conducted quantitative study. The discovered trends are the same for Western research. There is a visible slightly weaker protection of privacy by women than by men. At the same time, women are assigned higher care to not disclose such obvious elements as a telephone number or name and surname; however on the other hand, women reveal more of other data that can be deanonymised without being aware of it [13]. In the case of Polish society, this is true as long as we do not use the variable age as the control variable. We observe a stronger correlation between the lack of privacy competences and age than between the level of privacy protection and gender. We also know that in the over 55 years of age group, the females are overrepresented, so the differences can be explained in two ways; i.e., both age and gender. Differences in the perception of privacy as a value and age are widely recognized [34]. Moreover, in the literature on the subject, the differences between men and women are not taken as obvious [35].

The privacy paradox has also been confirmed to a limited extent for the Polish population. Susanne Barth and Menno D.T. de Jong identified three explanations for the privacy paradox [36] based on the analysed literature. First, the risks and benefits to privacy are rationally considered by actors, but the benefits outweigh the risks to privacy. Second, the threats and benefits to privacy are rationally weighed by actors; however, the result of the reasoning is distorted by irrational factors or limited rationality. On the other hand, the third explanation indicates that threats to privacy are not considered by users at all. The first and the second hypothesis seem to be the most probable. The first hypothesis is supported by the low intensity of

negative attitudes towards corporations in all surveyed groups in the entire Polish population of the country. The lack of systematic privacy protection behaviour, which is dictated by the lack of technical competences, speaks in favor of the second hypothesis. At the same time, the value of privacy is highly appreciated. The fact that there is a privacy paradox in Poland is supported by the correlation between education and the number of privacy protection measures used.

Relationships between political self-identification and attitudes towards privacy on the Internet are noticeable. In the moral aspect, the center-right and center-left are stronger than other political options against government surveillance. On the other hand, the right-wing highly value anonymity on the Internet and considers it unethical to limit it. The left-right, left, and right, to a large extent value all the elements of freedom in the moral aspect. In the legal aspect, the left and, to a slightly lesser extent, the right are against corporations that violate users' privacy. Behaviourally, the supporters of certain political options do not differ from each other. This diversity is not limited to Poland only [37].

5. Conclusions

In this article, we analysed the value system of Poles in terms of the phenomenon of privacy on the Internet. The following aspects were taken into account: privacy on the Internet as a moral value, privacy on the Internet as a subject of legal regulations (current or future), and actual actions taken by users to protect privacy. The differentiation of Polish society in terms of the three above-mentioned areas was also examined, creating its segmentation.

Freedom, including its component of privacy on the Internet, is valued by Poles. It is primarily verbal, because this value is, however, detached from reality and rarely translates into specific actions. There is a polarization of attitudes regarding privacy on the Internet as a subject of legal regulations. Poles are moderately willing to accept laws against tracking and profiling on the Internet. However, Poles remain significantly opposed to the legal regulation of the obligation to disclose their identity on the Internet and the legally guaranteed access to any information belonging to an ordinary user by law enforcement agencies. The basic observation is the low competence of Poles in the field of privacy protection on the Internet. Poles generally undertake ineffective forms of protection. The tools used are ineffective or mixed with ineffective or even counter effective (at most one in ten Polish users is able to effectively protect their privacy on the Internet).

A moderately strong, negative attitude towards state institutions was identified in the Polish society as a factor that could excessively limit privacy on the Internet. Although the negative attitude towards entities of the second sector prevails in Polish society, it is still lower than the negative attitude towards state institutions. Probable explanations of this phenomenon are a rational profit and loss calculation made by users (users receive benefits from Internet service providers, and therefore agree in exchange for tribute in the form of privacy violations) or insufficient knowledge about privacy violations by Internet service providers. Socio-demographic variables

differentiating Poles in terms of attitudes towards privacy on the Internet were identified. These attitudes depend mainly on: age, education, and (to a lesser extent) gender, socioeconomic status, and the size of the place of residence. The self-identification of political attitudes on the left-right scale is also a differentiating factor. However, the basic factor differentiating attitudes turned out to be the behavioural aspect, i.e., the actual actions taken by individuals to protect their privacy on the Internet. Low trust in state institutions in the field of privacy protection on the Internet may lead to a situation of negative attitudes towards all legal regulations and resistance to them. The issue of violations of privacy by Internet service providers, in particular social media, seems to be a dangerous and requiring awareness phenomenon; however, the lack of trust in state institutions may cause resistance to the reception of values conveyed by a potential social campaign. The level of the ability to protect privacy on the Internet by Poles is insufficient and far unsatisfactory, which may be important for maintaining the state of national security.

References

[1] LeakLookup, *Data Breach Search Engine*. [Online]. Available: https://leak-lookup.com/. [Accessed: Oct. 28, 2022].

[2] A. Florek, "The problems with PRISM: How a modern definition of privacy necessarily protects privacy interests in digital communications," *UIC John Marshall Journal of Information Technology & Privacy Law*, vol. 30, no. 3, pp. 571–606, 2014.

[3] L. Rainie, J. Anderson. (2014). *The Future of Privacy*. [Online]. Available: https://www.pewresearch.org/internet/2014/12/18/future-of-privacy/. [Accessed: Oct. 29, 2022].

[4] D. J. Solove, *Understanding privacy. Cambridge*, MA: Harvard University Press, 2008.

[5] H. Nissenbaum, *Privacy in context: technology, policy, and integrity of social life*. Stanford, CA: Stanford University Press, 2010.

[6] L. Rainie, S. Kiesler, R. Kang, M. Madden. (2013). *Anonimity, Privacy, and Security Online*. [Online]. Available: https://www.pewresearch.org/internet/2013/09/05/anonymity-privacy-and-security-online/. [Accessed: Oct. 29, 2022].

[7] Dell. (2022). *Dell Technologies Global Data Protection Index*. [Online]. Available: https://www.dell.com/en-sg/dt/data-protection/gdpi/index.htm. [Accessed: Oct. 29, 2022].

[8] Surfshark. (2021). *Wyniki ogólnopolskiego badania wskazują, że Polacy nie doceniają wartości swoich danych*. [Online]. Available: https://surfshark.com/pl/blog/polacy-nie-doceniaja-wartosci-swoich-danych. [Accessed: Oct. 29, 2022].

[9] M. Sadurski. (2022). *Wolność przede wszystkim, potem równość – to dla Polaków podstawowe wartości.* [Online]. Available: https://www.newsweek.pl/biznes/co-polacy-mysla-o-swoim-panstwie-wolnosc-przede-wszystkim-potem-rownosc/21hrn6q. [Available: Oct. 29, 2022].

[10] M. G. Hoy, G. Milne, "Gender Differences," in "Privacy-related measures for young adult Facebook users," *Journal of Interactive Advertising*, vol. 10, no. 2, pp. 28–45, 2010, doi: 10.1080/15252019.2010.10722168.

[11] K. Christopherson, "The positive and negative implications of anonymity in Internet social interactions: 'On the Internet, nobody knows you're a dog'," *Computers in Human Behavior*, vol. 23, no. 6, pp. 3038–3056, 2007, doi: 10.1016/j.chb.2006.09.001.

[12] K. B. Sheenan, "An investigation of gender differences in on-line privacy concerns and resultant behaviors," *Journal of Interactive Marketing*, vol. 13, no. 4, 1999, doi: 10.1002/(SICI)1520-6653(199923)13:4<24::AID-DIR3>3.0.CO;2-O.

[13] Y. Feng, W. Xie, "Teens' concern for privacy when using social networking sites: an analysis of socialization agents and relationships with privacy protecting behaviours," *Computers in Human Behavior*, vol. 33, pp. 153–162, 2014, doi: 10.1016/j.chb.2014.01.009.

[14] V. Steeves, P. Regan, "Young people online and the social value of privacy," *Journal of Information, Communication and Ethics in Society*, vol. 12, no. 4, pp. 298–313, 2014, doi: 10.1108/JICES-01-2014-0004.

[15] E.-M. Schomakers, Ch. Lidynia, L. Vervier, A. Gadeib, M. Ziefle, "Online privacy perceptions of older adults," *International Conference on Human Aspects of IT for the Aged Population*, 2017, doi: 10.1007/978-3-319-58536-9_16.

[16] A. Acquisti, „Privacy in electronic commerce and the economics of immediate gratification," *Proceedings of the 5th ACM Conference on Electronic Commerce*, 2004, pp. 21–29, doi: 10.1145/988772.988777.

[17] S. Barth, M. D. T. de Jong, M. Junger, "Lost in privacy? Online privacy from a cybersecurity expert perspective," *Telematics and Informatics*, vol. 68, 2022, doi:10.1016/j.tele.2022.101782.

[18] A. Deuker, "Addressing the privacy paradox by expanded privacy awareness: The example of context-aware services," in *Privacy and identity management for life*, M. Bezzi, P. Duquenoy, S. Fischer-Hüber, M. Hansen, G. Zhang, Eds. Berlin: Springer, vol. 320, 2010, pp. 275–283.

[19] Z. Benenson, O. Kroll-Peters, M. Krupp, "Attitudes to IT security when using a smartphone," *Proceedings of the Federated Conference on Computer Science and Information Systems*, 2012, pp. 1179–1183.

[20] I. Shklovski, S. D. Mainwaring, H. H. Skúladóttir, H. Borgthorsson, "Leakiness and creepiness in app space: Perceptions of privacy and mobile app use," *Proceedings of the Conference on Human Factors in Computing Systems*, 2014, pp. 2347–2356, doi: 10.1145/2556288.2557421.

[21] S. Barth, M. D. T. De Jong, M. Junger, P.H. Hartel, J.C. Roppelt, "Putting the privacy paradox to the test. Online privacy and security behaviors among users with technical knowledge, privacy awareness, and financial resources," *Telematics and Informatics*, vol. 41, pp. 55–69, 2019, doi: 10.1016/j.tele.2019.03.003.

[22] S. D. Warren, L. D. Brandeis, "The right to privacy," *Harvard Law Review*, no. 5, pp. 193–220, 1890.

[23] Y. Tsukada, K. Mano, H. Sakurada, Y. Kawabe, "Anonymity, privacy, onymity, and identity: A modal logic approach," *Transactions on Data Privacy*, no. 39, pp. 177–198, 2010.

[24] J. Assange, J. Appelbaum, M.-M.J. Zimmermann, *Cypherpunks: Freedom and the future of the Internet*. New York, London: OR Books, 2016.

[25] J. Waksberg, "Sampling methods for random digit dialling," *Journal of the American Statistical Association*, vol. 73, pp. 40–46, 1973.

[26] R. F. Potthoff, "Some generalisation of the Mitofsky-Waksberg technique for Random Digit Dialling," *Journal of the American Statistical Association*, vol. 82, pp. 409–418, 1982.

[27] H. E. Driver, A. L. Kroeber, "Quantitative expression of cultural relationships," *University of California Publications in Amer. Archaeology*, vol. 31, pp. 211–256, 1932.

[28] J. Czekanowski, "Objectiv kriterien in der ethnologie," *Korrespondenzblatt der Deutschen Gesselschaft fur Anthropologie, Ethnologie, und Urgeschichte*, vol. 47, pp. 1–5, 1911.

[29] R. B. Cattell, "The description of personality: Basic traits resolved into clusters," *Journal of Abnormal and Social Psychology*, vol. 38, pp. 476–506, 1943, doi:10.1037/h0054116.

[30] R. R. Sokal, P. H. Sneath, *Principles of numerical taxonomy*. San Francisco-London: Freeman 1963.

[31] R. K. Blashfield, "The growth of cluster analysis: Tryon, ward, and johnson," *Multivariate Behavioral Research*, vol. 15, no. 4, pp. 439–458, 1980.

[32] CBOS. (2020). *Wartości w czasach zarazy, no. 160*. [Online]. Available: https://www.cbos.pl/SPISKOM.POL/2020/K_160_20.PDF. [Accessed: Oct. 30, 2022].

[33] CBOS. (2021). *Młodzi Polacy o zasadach demokracji, no. 120.* [Online]. Available: https://www.cbos.pl/SPISKOM.POL/2021/K_120_21.PDF. [Accessed: Oct. 30, 2022].

[34] N. Demertzis, K. Mandenaki, Ch. Tsekeris, "Privacy attitudes and behaviors in the age of post-privacy: An empirical approach," *Journal of Digital Social Research*, vol. 3 no. 1, pp. 119–152, 2021, doi:10.33621/jdsr.v3i1.75.

[35] T. Dienlin, S. Trepte, "Is the privacy paradox a relic of the past? An in-depth analysis of privacy attitudes and privacy behaviours," *European Journal of Social Psychology,* vol. 45, no. 3, pp. 285–297, 2015.

[36] S. Barth, M.D.T. de Jong, "The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review," *Telematics and Informatics*, vol. 34, no. 7, pp. 1038–1058, 2022, doi: 10.1016/j.tele.2017.04.013.

[37] C. Neill. (2021). *Politics of Privacy: The Role of Individual Political Views in Consumer Data Privacy Concerns, Honor Theses*. [Online]. Available: https://egrove.olemiss.edu/hon_thesis/1697. [Accessed: Oct. 30, 2022].

The Substantive Criminal Aspects of the Offence of Simulated Child Pornography under Polish Law

Remigiusz Rosicki | **Adam Mickiewicz University in Poznan, Poland**,
ORCID: 0000-0002-1187-5895

Abstract

The objective scope of the research problem concerns the content and sense of the elements characterising one of the types of child pornography, criminalised under Art. 202 §4b of Poland's Criminal Code, i.e. simulated child pornography. This offence is understood as producing, disseminating, presenting, storing and possessing pornographic material presenting a generated or processed image of a minor participating in sexual activity. The main goal of the article is the substantive criminal analysis of the act criminalised under Art. 202 §4b of the Code. The scope of the analysis has been elaborated with the following question: To what degree is the legal solution concerned with criminalisation and penalisation of the activities of »production, dissemination, presentation, storage or possession of pornographic material presenting a generated or processed image of a minor participating in sexual activity« usable, and realises the ratio legis intended by the legislator? The issue has been analysed using what is primarily an institutional and legal approach, involving textual, functional and doctrinal interpretations that have been supplemented by the author's own conclusions and opinions.

Keywords
pornography, child pornography, sexual offences, information restriction, cybercrime

Corresponding author:
Remigiusz Rosicki, Adam Mickiewicz University in Poznan, ul. Wieniawskiego 1, 61–712 Poznan, Poland;
ORCID: 0000-0002-1187-5895;
remigiusz.rosicki@amu.edu.pl

Cite this article as: R. Rosicki, "The Substantive Criminal Aspects of the Offence of Simulated Child Pornography under Polish Law", vol. 1, no. 1, pp. 272–284, 2022, DOI: 10.5604/01.3001.0016.0690

1. Introduction

The material scope of the research problem in the text encompasses selected issues concerning the content and sense of the elements characterising one of the types of child-pornography offence criminalised under Art. 202 of the Polish Criminal Code. The object of in-depth analysis is the crime of simulated child pornography, which is understood to be material presenting a generated or processed image of a minor participating in sexual activity (Art. 202 §4b of the Criminal Code). The crime of simulated child pornography was criminalised through an amendment of the Criminal Code dated 24 October 2008 [1]. It is accepted that the ratio legis behind the criminalisation of this act is the intention to eliminate pornographic material involving an image of a minor under the age of 18. This fits with solutions arrived at in international law with a view to various types of child pornography being criminalised and penalised. The relevant (2007) Council of Europe Convention signed at Lanzarote, and concerned with the protection of children against sexual exploitation and sexual abuse, already introduced a definition of child pornography – and alongside imagery of a child participating in a real sexual act, it also makes reference to participation in simulated sexual acts [2]. Following on from that, the EU's 2011 Directive of the European Parliament and of the Council regards as child pornography "any material that visually depicts any person appearing to be a child engaged in real or simulated sexually explicit conduct or any depiction of the sexual organs of any person appearing to be a child, for primarily sexual purposes", as well as "realistic images of a child engaged in sexually explicit conduct or realistic images of the sexual organs of a child" [3].

Against that background, this paper has as its main purpose the substantive crime-related analysis of the offence of simulated child pornography, as in turn criminalised in Poland under Art. 202 §4b of the Criminal Code; as well as evaluation of this legal solution arrived at in respect of its usefulness on the one hand, and its criminal function on the other. With a view to this research topic being further elaborated on, the research question posed was: To what degree is the legal solution concerned with criminalisation and penalisation of the activities of »production, dissemination, presentation, storage or possession of pornographic material presenting a generated or processed image of a minor participating in sexual activity« usable, and realises the ratio legis intended by the legislator?

The analysis referred to is primarily an overview, mainly from the institutional and legal points of view, with the aid of textual, functional and doctrinal interpretations. However, the effects have also been supplemented by the author's own conclusions and opinions as regards de lege lata solutions.

2. Models for the criminalisation of pornography

While the term "pornography" is derived from the two Greek words for "prostitute" (πρνη, porne) and for "write" (Γραφός, graphos), the term in the sense we use it today would not have been known to the ancients. we use

it today would not have been known to the ancients. Indeed, Athenaeus of Naucratis in the 3rd century CE wrote one of the first works featuring the term πορνογράφον (Greek, pornográfon), which was *Deipnosophistae* ("The Dinner Sophists"). It was not then until the 19th century that our present word 'pornography' came into widespread usage. However, as of that time it was still literary or scientific work devoted to the phenomenon of prostitution that was being referred to [4].

Nevertheless, by the 1890s, there was already both theoretical discussion and practical work as regards legal solutions that might achieve the elimination of pornography from the public sphere. Filar recalls how issues resolving around the counteraction of immoral content were the subject of the international convention and conference done in Paris in 1908 and 1910. An international accord was thus arrived at as regards the prevention of the circulation of pornographic publications, with the first regulation in international law in this field being the Geneva Convention of 1923, whose purpose was to combat the circulation of and trade in pornographic publications [5, p. 25, 6, pp. 43–58].

However, the 1923 Convention offered no legal definition of its core subject and term, indeed emphasising that it would be up to individual State Parties to decide on how to define what was 'pornographic' (of obscene as the French had it). What the Convention did do was characterise pornography by reference to material carriers thereof, be these writings, pictures, drawings, paintings, prints, images, posters, emblems, photographs, films or other objects [7].

Individual countries go on regulating (criminalising or decriminalising) the phenomenon of pornography, with this arising *inter alia* out of different ways in which such criminalisation or decriminalisation of pornographic media are actually achieved. It is cultural differences as well as different legal systems in states that stand in the way of a uniform position being developed by the international community. Equally, whatever the detailed circumstances, we are left in doubt that what is involved here is an attempt at regulating private life with the aid of instruments relating to the public sphere. This leaves actions taken as seeking to give effect to a specific legal policy in the criminal sphere. Generalising from that we note how criminal policy encompasses a set of actions a state takes to combat crime as such, but also to eliminate negative phenomena in society often constituting a prelude to the perpetration of the crime(s) in question. There is then penal policy, as one element of criminal policy, that takes in actions by which: (1) penal law is laid down, (2) penal law is applied, (3) laid-down and applied law is assessed from the point of view of its effects, and (4) the effectiveness of laid-down and applied law is in turn assessed [8, pp. 229–265, 9, pp. 122–190, 10, pp. 11–43, 11, pp. 123–141].

Against that background it is then possible to point to various models of criminal policy relating to pornography, as these may be: (1) restrictive, (2) restrictive-permissive, and (3) permissive. The first model provides for complete prohibition of pornography, while taking into account different categories thereof present in different legal systems. The second model is one involving mixed strategies of criminal policy, which bring about the effect of either legal relevance or irrelevance of pornographic

material. More often than not, this results from division of pornographic material into harmful and harmless content. In accordance with the third model, pornographic material is not criminalised, and so it is irrelevant to penal policy. It seems that the model applied most frequently is the restrictive-permissive one, which in certain cases does not criminalise pornographic material, while singling out particular types of pornography that are subject to strict prohibition.

A visible process in the criminal policy pursued by countries is one whereby specific types of pornographic material are criminalised, e.g. paedophilia, or the use or presentation of violence. This may in fact be exemplified by the criminalisation of simulated child pornography, which is the object of analysis here. Equally, it is reasonable to conclude that the third of the above-mentioned models of criminal policy as regards pornography is rather uncommon these days.

3. The concept and definitions of pornography

The Polish legislator has neither presented a legal definition of pornography nor used an indicium thereof, though pornographic content has been addressed. This has perforce left the interpretation of the indicium in question to the doctrine of penal law. To understand what is pornographic in nature, the literature has put forth various ways for the category to be demarcated. Thus the aforesaid Filar presents conditions needing to be met concomitantly for given material to be deemed pornographic. The author points out that material of this kind should: (1) contain sexual acts that differ substantially from accepted societal standards in this regard, (2) focus solely on the technical aspects of sex life and sexuality itself, while dehumanising sexual acts by presenting only their sexual functions, as well as the gynaecological and anatomical structure of sex organs, (3) be verified with regard to the author's motivation, e.g. willingness to show technical aspects of sexual acts or cause sexual arousal in viewers, (4) be verified as to technical and aesthetic level, where the premise is that the lower the level, the greater the likelihood that a given work will be regarded as pornography [12, pp. 39–40, 13, pp. 202–206].

It may be noted how the above-mentioned characteristics of pornographic material appear to give too much competence to the representatives of the doctrine with regard to the evaluation of material potentially constituting pornographic material. They are highly dubious, as certain material might conceivably feature pornographic content, even as it fails to meet most of the conditions referred to. By way of illustration, while Lars von Trier's Antichrist is not a pornographic film, certain of its scenes may be considered pornographic. This leaves it possible to distinguish between something referred to in its entirety as a pornographic film and something that is referred to as pornographic material within the meaning of the doctrine of penal law. This results from the display of sex organs involved in sexual acts being a core characteristic of pornography. Equally, what conventionally needs to be conveyed is such an involvement of the said organs as shows them directly, rather than metaphorically, implicitly or by means of manifold hiding filters.

In addition, attention should be drawn to the thesis regarding the interpretation of the category of pornographic material that the Supreme Court of the Republic of Poland presented in its ruling of 2010. It was held that the category of pornographic material was to be regarded as a legal, rather than a medical or sexological, category. This leaves experts competent solely to assess the potential effect specific material may have on a viewer when it comes to reactions (in essence of a sexual nature). Proceeding from that, no expert (even a sexologist) can stand in for a court of law when it comes to assessing whether the indicium for pornographic material under given circumstances has or has not been fulfilled [14, 15, pp. 128–151].

It follows from the above comments on the characteristics used in demarcating the boundaries of pornographic material that it is necessary to rely on the category of sexual act. Acts of this type include both sexual relations and other sexual acts. Both types of act constitute indicia of the offence of rape criminalised under Art. 197 of the Criminal Code.

Sexual relations are understood, in the first place, as sexual intercourse (coitus), which involves the insertion of male genitalia into female genitalia. However, it is noteworthy that the category of sexual relations is broader than the category of a sexual intercourse. The scope of sexual relations also includes surrogates for sexual intercourses, i.e. any forms of sexual contact regarded as comparable or equivalent thereto. It is worth pointing out that the doctrine features variant interpretations as to what can be termed a "surrogate for a sexual intercourse." Filar regarded as such surrogates direct sexual contact involving body parts of one of the sexual-act participants and body parts of the other participant, which, despite not being actually sexual, are considered comparable to sex parts by the perpetrator, who uses them to satisfy his or her sexual needs. To Warylewski, surrogates for sexual intercourses are those in which the necessary element is the involvement of either the perpetrator's or the victim's sex organs. The author referred to found this element necessary, but not sufficient, considering that surrogates for sexual intercourses are confined to penetrations of natural body orifices in the circumstances of copulation being imitated. This provides for the inclusion as types of sexual intercourse of oral sex in the form of cunnilingus, axillary and interfemoral intercourse, and coitus vestibularis. To Rodzyniewicz, surrogates for sexual intercourses are those acts that involve direct contact between the perpetrator's body and the victim's sex organs, or other body parts of him or her, that the perpetrator finds equivalent and sexually satisfying [16, pp. 211–229, 17, 13, pp. 45–59].

The jurisprudence has come to feature various interpretations of acts as sexual relations and surrogates for them, which comprise: (1) oral sex, (2) the insertion of a hand into a female reproductive tract, (3) the insertion of a penis-shaped vibrator into a victim's vagina and anus, (3) the insertion of a bottleneck into the anus (but also other objects, e.g. a stick). From this it can be seen readily enough how the jurisprudence extends beyond the doctrinal definitions of sexual relations. For instance, when the perpetrator penetrates the victim's body with inanimate objects, that may be regarded as surrogates for male genitalia.

Another element in the category of sexual acts is another sexual act. Of course, the scope of other sexual acts does not cover sexual intercourse. Still, other sexual acts are connected with sexuality *sensu largo*. As regards the fulfilment of indicia, they denote carnal relations involving intimate parts, without sexual intercourse, but with a culturally sexual character attributed to them, i.e. the perpetrator's sexual arousal. It is accepted that intimate parts include genital, vaginal, anal and mammary-gland areas. Examples of other sexual acts include the masturbating of a victim, as well as instances of a victim being forces to masturbate or to masturbate another person, touching genital, vaginal and mammary-gland areas [16, pp. 211–229, 18, pp. 132–194, 19, 20]

Given the singling out of the category of a surrogate for sexual acts, and the attempt at explicating its scope, the borderline between sexual intercourse and another sexual act becomes shifted in both doctrine and jurisprudence. Various stances in this respect can be exemplified by legal classifications concerning the insertion by the perpetrator of objects, which are not elements of his or her body, into the final segment of the digestive tract or of the female reproductive system. Other examples include acts of masturbation, which by some commentators would be reckoned among other sexual acts, while by others among surrogates for sexual intercourse, and hence sexual relations [18, pp. 132–194].

Alongside the category of pornographic material in the penal provisions, the Polish legislator uses the indicium of the image of a naked person (see Art. 191a of the Criminal Code – *the offence of recording and distributing the image of a naked person or a person during sexual activity*). Following the doctrine, the image of a naked person means exposed intimate parts, e.g. sex organs and mammary glands. The image of a naked person is understood, not only as a stark-naked person, but also by reference to exposed intimate parts of him or her. At the same time, it is accepted that the image of a stark-naked person who has his or her intimate parts covered, e.g. by edited blackening, does not fulfil the indicium of the image of a naked person. The status of intimate parts not related directly to a person's face remains problematic. Still, it is accepted that if another characteristic element of a person's appearance allows him or her to be identified, then an image within the meaning of Art. 191a of the Criminal Code is deemed to be involved [21, pp. 567–571, 22, pp. 1023–1026]. Material containing images of naked persons or images of persons during sexual activity, within the meaning of Art. 191a of the Criminal Code, may become pornographic within the meaning of Art. 202 of the Criminal Code, if these bear characteristics discussed above with regard to the concept of pornography.

4. The offence of simulated child pornography

Simulated child pornography was criminalised in Poland as a result of the amendment of the Criminal Code in 2008. The criminalisation was intended to implement the Council Framework Decision 2004/68/JHA of 22 December 2003, which was no longer in force, in respect of the combating of the sexual exploitation of children, and child pornography. Pursuant to

the solutions contained in the Decision, the EU Member States were obliged to combat, not only pornographic material involving real children, but also material presenting individuals resembling children, or presenting realistic images of non-existing children [1, 22, pp. 1080–1086, 23].

In Art. 202 §4b, as the legislator uses the indicium of a generated or processed image of a minor, it is worth explaining here the very category of such an image. According to a dictionary definition, an image means a depiction, likeness, portrait and representation of something, also a photograph (the word 'image' has its Polish equivalent of 'wizerunek', which is derived from the German 'Viserung'). Besides, the word 'image' can be synonymised with such words as reflection, mirroring and reconstruction. Hence, in the early Polish language the actions of regarding something attentively, targeting something with your eye or measuring were referred to as 'wizerowanie' [24, p. 1233, 25, pp. 342–343, 26, p. 1172].

Doctrinally, a generated or processed image of a minor has come to be referred to as simulated child pornography. And so such adjectival indicia as "generated" and "processed" have come to be identified with another one – simulated. According to a dictionary definition, the adjectival expression "processed" means 'creatively transformed,' 'changed,' 'with a different shape or appearance'. According to a dictionary definition, the adjectival expression "generated" means 'produced,' 'crafted,' 'created,' 'caused by something' or 'made.'

An image is to be understood as the likeness of a specific person, which allows him or her to be identified. The features enabling identification of a person are physical, and so they are connected with appearance. For a person's image to be identified, it is not necessary for his or her face to be presented, as other body parts may serve that purpose. However, it is noteworthy that, in pornographic material, a person's face is usually used in connection with other parts of the body. Still, the Polish legislator uses the indicium of the image of a minor. This means that the image of a person must be identifiable as the image of a minor. At the same time, the legislator does not indicate the criteria by which minority characteristics are to be classified – which would not be problematic in the case of "ordinary" child pornography (where the real age of a minor is a firm criterion), but it can still cause problems when it comes to indicating characteristics of a generated or processed image of a minor. A relatively clear example – one that illustrates the fulfilment of the indicium of the *ratione materiae* under Art. 202 §4b – may entail intentional generation of animation with a plot allowing for no other possible interpretation but to accept that it presents a generated image of a minor.

The above does not preclude the presentation of a quite problematic example whereby production of a pornographic film involves a major with an image (appearing or stylised in a manner) imitating the appearance of a minor, or which involves sex dolls imitating an image of a minor.

It needs to be borne in mind that a minor's naked image alone does not suffice, as it needs to be related to sexual acts, and only then does it amount to pornographic material. Next to the previously presented doctrinal interpretations of the indicia of sexual activity (sexual intercourse and other sexual acts), it is worth mentioning solutions present in the Council Framework Decision 2004/68/JHA of 2003. The latter defines

simulated child pornography as: (1) realistic images of a non-existing child participating in clearly sexual activity, or being subjected to such activity, including lewd presentation of sex organs or intimate parts, (2) images of a real person who appears to be a child, participating in clearly sexual activity, or being subjected to such activity, including lewd presentation of sex organs or intimate parts [23].

As regards the causative acts criminalised under Art. 202 §4b, the legislator includes: (1) producing, (2) disseminating, (3) presenting, (4) storing, and (5) possessing pornographic material presenting generated or processed images of minors participating in sexual activity. Interestingly enough, compared with §3, in §4 the legislator does not include such causative acts as recording and importing. Compared with the Sections in Art. 202 of the Criminal Code, which do not mention a direct intent as an action aimed at dissemination, as an indicium of the subjective side, i.e. §4 and 4a, the legislator has not included such indicia as recording and gaining access. Thus, one can point to a lack of coherence with regard to the regulations criminalising the offence of pornography, or even to a lack of logic. This results from the fact that the legislator has excluded recording, importing and gaining access from the verbal features of the indicia of simulated child pornography. Hence, the objective scope of the criminalisation of the offence of simulated child pornography does not include, by way of illustration, a situation in which an Internet user gains access, while browsing, to pornographic material bearing the indicia specified in Art. 202 §4b of the Criminal Code. At the same time, if the perpetrator has done that in connection with pornographic material involving a minor in general, the act would be punishable.

The verbal feature "produces" denotes a whole process of producing pornographic material for specific purposes, taking into account technical and organisational aspects. The most extreme examples of the fulfilment of this indicium are acts performed by producers, directors, scriptwriters, operators and even actors. And so it follows that the indicium of production of pornographic material is fulfilled by every participant in the organisational and technical activities undertaken as necessary in the process [22, pp. 1080–1086].

Production of animated films presenting minors participating in sexual activity may be considered to fulfil the indicia specified in Art. 202 §4b of the Criminal Code. Generated by means of a variety of methods, characters in such cartoons ought to be recognised as fulfilling the indicium of the generated image of a minor. Of course, a minor engaged in sexual activity needs to be presented in a specified manner for given film material to be regarded as pornographic. Theoretically, the use of a minor's generated image may be related to a specific type of *hentai* films and comics, e.g. *rorikon* (material most commonly presenting pre-pubescent girls), *shotacon* (material most commonly presenting pre-pubescent boys), *ecchi* (material presenting sexual fantasies, most often involving sexually ignorant attractive women, which sometimes takes the form of a playful plot), *yuri* (material presenting lesbian relationships), *yaoi* (material presenting gay relationships with dominant and subdued roles), *kinbaku* (material showing the tying and restraining of bodies, e.g. bondage), *tenacle* or *shokushu-kei* (material presenting women's sexual

activity involving tentacled non-human creatures), *bukkake* (material presenting sexual activity in which men ejaculate on women, and most often their faces), *gokkun* (material presenting sexual activity in which women swallow sperm), *futanari* (material presenting sexual activity involving outstanding sex organs, frequently featuring group sex and *bukkake*) [27, pp. 340–353, 28].

It is also noteworthy that, irrespective of the strict use of the category of hentai, in the sexual context, in the pornography industry, a category that draws freely on the convention of this type of animated or comic material has been singled out. As a result, on pornographic websites, tabs with such keywords are created, but they cover both animated material and films featuring actors in the industry.

The problem concerned with the production of pornographic material with generated or processed images of minors may grow in strength in connection with the development of *deepfake* technology. The availability of software offering possibilities for creating and processing personal images gives rise to twofold problems. The first problem is the greater harmfulness of this type of acts to the person who has fallen victim to, by way of illustration, dissemination of pornographic material containing his or her image. The other problem is perpetrators' lack of awareness of criminalisation of this type of conduct, or the sense of anonymity resulting from the universal availability of the Internet [29, pp. 133–140, 30, pp. 39–52]. Frequently, this phenomenon is related to some forms of stalking, insult or slander. Arguably, every use of an image of a minor, where it is pasted in a scene of some existing pornographic material involving majors, or *vice versa*, fulfils the indicia of generating and processing, which are mentioned in Art. 202 §4b of the Criminal Code. Incidentally, it is worth stressing the problem concerned with the scale of interference in material, and of the use of, for instance, an existing image of a minor, which is by no means specified with indicia, in fact.

The indicium of dissemination in the general sense means making generally known, available and public. This means that dissemination of pornographic material is about making it available to a larger and/or unlimited number of people. Such an action may be taken, for instance, by putting such material on the Internet (e.g. on someone's own websites, open accounts of social messaging services, accounts of pornography website users). It is therefore accepted that acts consisting in making pornographic material available to a small number of persons, or to a circle of persons defined narrowly or strictly does not fulfil the indicium of dissemination. It is to be noted that the indicium of dissemination is not equivalent to another one, that of publicising. This means that dissemination of pornographic material does not have to be public. Still, it has been accepted in the jurisprudence that lending, copying, publicising and other forms by which material of the above type is made available to an unspecified number of persons is equivalent to dissemination [21, pp. 665– 676, 22, pp. 1080–1086]. Arguably, the putting on the Internet of pornographic content with a minor's image that has been generated or processed by: (1) pasting a minor's face to a photograph such that it replaces the face of a major participating in sexual activity, (2) creating *deepfake* where a face of a major participating in sexual activity is replaced with a face of a minor,

(3) creating a realistic animated film showing sexual acts performed by minors – all instantiate the fulfilment of the indicia of dissemination, as referred to in Art. 202 §4b of the Criminal Code.

The indicium of presentation is understood as the showing of pornographic material subject to criminalisation. Presentation can also be effected through various forms of media, e.g. films, magazines, pictures, photographs and posters; and it can be done both publicly and non-publicly. Still, it should be noted that presentation of pornographic material by means of the Internet is often public, given the character of that medium, and hence involves the fulfilment of the indicium of dissemination. In a case like this, the active or passive attitude of the viewer of pornographic material is of no consequence. This in turn means that for the indicium of presentation to be fulfilled, it does not matter whether the viewer of criminalised material took the initiative in becoming exposed to it. However, it is accepted that, in a situation where becoming exposed to pornographic material requires the viewer's conscious involvement, there is no fulfilment of the indicium of presentation [13, p. 208, 21, pp. 665–676, 22, pp. 1080–1086, 31]. As regards pornographic material showing a generated or processed image of a minor, the fulfilment of the indicium of presentation will be the case if the perpetrator for instance uses e-mail to make available videos and photographs containing such material. The perpetrator may also present criminalised material by making it available on social media accounts (e.g. *Facebook*).

Storage is associated with the holding of pornographic material, which is most often done clandestinely. The legislator does not point to the timescale for storage of this type of pornographic material. However, alongside storage, the legislator has included the indicium of possession of pornographic material, which is to be understood as factual control over it (autonomous or dependent possession within the meaning of the civil law). Hence, having control over pornographic material on behalf – or for the sake – of other persons fulfils the indicium of storage, and not possession. Still, it is noteworthy that possession is the case after, for instance, a file containing pornographic material showing simulated child pornography has been downloaded and recorded on a data storage device; and when such material is stored in the so-called "Cloud", to which there may be unlimited access. This assumption follows from the fact that the scope of criminalisation does not cover the data-storage device itself, but the pornographic material [21, pp. 665–676, 22, pp. 1080–1086, 32; 33].

5. Conclusion

The object of analysis in the text is the content and sense of the indicia of the offence of so-called simulated child pornography, as criminalised under Art. 202 §4b of the Polish Criminal Code. The said analysis, which is of a substantive criminal nature, resorts to textual, functional and doctrinal interpretations, albeit with outcomes supplemented by the author's own conclusions and opinions. With a view to the material scope of the analysis being elaborated, and conclusions presented, the formulated research question was: To what degree is the legal solution concerned with criminalisation

and penalisation of the activities of »production, dissemination, presentation, storage or possession of pornographic material presenting a generated or processed image of a minor participating in sexual activity« usable, and realises the ratio legis intended by the legislator?

The rationale behind the criminalisation and penalisation of simulated child pornography under Polish criminal law has been to eliminate harmful effect on the personality of a viewer. Seen from a broader perspective, an imaginary (artificially created or processed) exploitation of minors would be conducive to the culture of permissiveness with regard to sexual violence directed at this group. The issue of simulated child pornography was addressed in some degree by the Council of Europe Convention done at Lanzarote, which also included, alongside the imagery of a child participating in real sexual activity, participation in simulated sexual acts. At the same time, the Convention singled out the category of simulated presentation or realistic images of a non-existing child.

The textual analysis demonstrates the ambiguity of some of the indicia intended to characterise the category of simulated child pornography. While the terms 'generation' and 'processing' are understandable linguistically, as has to be conceded, assignment of factual creations to them may prove problematic. This reflects a lack of standards as regards generation and processing, their effects or assessment of the age of a person in generated imagery. Nor has the legislator indicated the degree of processing of a minor's image, which means that it is unclear what criterion for the realism of a minor's image should be adopted. Additionally, criminalisation of this type of pornography does not distinguish between a collage made by the perpetrator from photographs clipped manually, and video material made by the perpetrator by way of *deepfake* technology. Besides, legal classification of simulated child pornography is problematic, because the process could involve design in the virtual worlds of games like *Second Life*.

Another problematic aspect is the stage at which an image or creation of simulated sexual acts is made subject to processing. This can be seen clearly in the process of selection for the purposes of pornographic films such actors as have juvenile traits, or can be made to look juvenile with the aid of make-up. This reveals a problem concerned with the difference between the creation of a minor's image with animated video technology and selection of suitable actors and make-up tricks. A similar situation will be one involving sex mannequins with juvenile traits (the so-called sex dolls), as these feature in video material or pornographic material of some other type. There is no doubt that, viewed objectively, there is no difference between the creation of a minor's image with the aid of animation, and the creation of such an image in the form of a sex mannequin used for sexual activity in pornographic material.

Given technological advancement (e.g. artificial intelligence in connection with *deepfake*) and the potential greater harmfulness of pornographic material, it is to be assumed that criminalisation of child pornography will not solely serve to eliminate deleterious effects on viewers, as well as specific content from social circulation. Rather, with the passage of time, it may also prove useful in eliminating such other acts as stalking, and the recording and dissemination of the image of a naked person participating in sexual activity.

References

[1] The Act amending the Criminal Code Act and certain other Acts (the Dziennik Ustaw Official Journal of Laws RP of 2008, no. 214, item 1344).

[2] "The Council of Europe Convention on the Protection of Children Against Sexual Exploitation and Sexual Abuse", Lanzarote, 25th October 2007, Dziennik Ustaw Official Journal of Laws RP of 2015, item 608.

[3] Directive 2011/92/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA.

[4] L.M. Nijakowski, *Pornografia, historia, znaczenie, gatunki*. Warsaw: Iskry, 2010.

[5] M. Filar, Pornografia. *Studium z dziedziny polityki kryminalnej*. Toruń: UMK, 1977.

[6] A. Przyborowska-Klimczak, "Zapobieganie i zwalczanie pornografii w świetle dokumentów międzynarodowych," in *Pornografia*, M. Mozgawa, Ed. Warsaw: Wolters Kluwer, 2011, pp. 43–58.

[7] International Convention for the Suppression of the Circulation of and Traffic in Obscene Publications, Geneva, 12th September 1923, Dziennik Ustaw Official Journal of Laws RP of 1927, no. 71, item 621.

[8] J. Lande, "Socjologia Petrażyckiego," *Przegląd Socjologiczny*, vol. 12, pp. 229–265, 1958.

[9] A. Podgórecki, Socjologia prawa, Warsaw: WP, 1962.

[10] B. Stańdo-Kawecka, *Polityka karna i penitencjarna między punitywizmem i menedżeryzmem*. Warsaw: Wolters Kluwer, 2020.

[11] Z. Ziembiński, *Socjologia prawa jako nauka prawna*. Warsaw-Poznań: PWN, 1975.

[12] M. Filar, *Przestępstwa seksualne w nowym kodeksie karnym*, in N*owa kodyfikacja karna. Kodeks karny*, Warsaw: Ministerstwo Sprawiedliwości, 1997, pp. 7–64.

[13] J. Warylewski, *Przestępstwa przeciwko wolności seksualnej i obyczajności.* Rozdział XXV Kodeksu karnego. Komentarz. Warsaw: Beck, 2001.

[14] Judgment of the Supreme Court of the Republic of Poland of 23 November 2010 (IV KK 173/10).

[15] W. Kozielewicz, „Przestępstwo produkowania i rozpowszechniania pornografii – kwestie sporne w świetle orzecznictwa Sądu Najwyższego," in *Pornografia*, M. Mozgawa, Ed., Warsaw: Wolters Kluwer, 2011, pp. 128–151.

[16] M. Bielski, "Wykładnia znamion „obcowanie płciowe" i „inna czynność seksualna" w doktrynie i orzecznictwie sądowym," *Czasopismo Prawa Karnego i Nauk Penalnych*, vol. 12, no. 1, pp. 211–229, 2008.

[17] Judgment of the Appeals Court in Katowice of 9 November 2006 (II AKa 323/06).

[18] M. Budyn-Kulik, "Inna czynność seksualna. Analiza dogmatyczna i praktyka ścigania," *Prawo w działaniu*, vol. 5, 2008, pp. 132–194.

[19] Decision of the Supreme Court of the Republic of Poland of 21 May 2008 (V KK 139/08).

[20] Judgment of the Supreme Court of the Republic of Poland of 5 April 2005 (III KK 187/04).

[21] J. Giezek, Ed., *Kodeks karny. Część szczególna. Komentarz*. Warsaw: Wolters Kluwer, 2021.

[22] A. Grześkowiak, K. Wiak, *Kodeks karny. Komentarz*. Warsaw: C.H. Beck, 2019.

[23] Council Framework Decision 2004/68/JHA of 22 December 2003 on combating the sexual exploitation of children and child pornography Official Journal L 013, 20/01/2004.

[24] B. Dunaj, Ed., *Słownik współczesnego języka polskiego*. Warsaw: Wilga, 1996.

[25] M.S.B. Lange, *Słownik języka polskiego*. Tom szósty. Lwów: Zakład Ossolińskich, 1860.

[26] J. Sobol, Ed., *Słownik wyrazów obcych*. Warsaw: PWN, 1995.

[27] T. Burdzik, "Hentai: erotyka z mangi i anime". *Kultura i Edukacja*, vol. 3, no. 103, pp. 340–353, 2020.

[28] M. McLelland, "A Short History of 'Hentai'," *Intersections: Gender, History and Culture in*

the Asian Context, vol. 12, 2006 [Online]. Available: http://intersections.anu.edu.au/issue12/mclelland.html. [Accessed: Jul. 25, 2022].

[29] C. Öhman, "Introducing the pervert's dilemma: a contribution to the critique of Deepfake Pornography *Ethics and Information Technology*, vol. 22, pp. 133–140, 2020.

[30] M. Westerlund, "The Emergence of Deepfake Technology: A Review". *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, 2019.

[31] Judgment of the Supreme Court of the Republic of Poland of 11 January 2017 (III KK 188/16).

[32] Judgment of the Supreme Court of the Republic of Poland of 17 May 2017 (III KK 478/16).

[33] Atenajos. "Deipnosophistae" [Online]. Available: http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A2008.01.0405%3Acasaubonpage%3D567b. [Accessed: Jul. 25, 2022].

UK Border Digitalisation – a Commentary on the Current State of Affairs

Marika Kosiel-Pająk | **Women In International Security, Poland**, ORCID: 0000-0002-6455-2133

Abstract

The commentary focuses on the current process of converting the British immigration procedures into an entirely digital format, as part of a reform brought about by Brexit and in the framework of broader digital strategies in the United Kingdom of Great Britain and Northern Ireland. The British government's ambitious aim is to digitalise the immigration procedures by 2025, further support eGates and eventually enforce a contactless mode of arrival. The policy plan, its execution to date and its reception are analysed briefly. Taking into account that the government is revealing only selected aspects of the complex system rather than all the mechanisms and safeguards, neither British digital sovereignty in this matter nor the scope of protection of personal and meta-data could be fully examined.

The challenges already encountered are set out, with the conclusion being that a human-centred approach is still lacking in the practical employment of the policy. Subsequently, the isolationist or populist concept of protecting the state against any migrant, as a potential threat rather than a potential to be developed for the benefit of the state, is the prevailing mindset. Finally, the current political and economic instability may play a pivotal role in policy implementation and contribute to its ultimate failure.

Keywords

border management, Brexit, contactless border, digital border, digital status, eGates, electronic travel authorisation, European Union Settled Scheme, immigration procedures, United Kingdom of Great Britain and Northern Ireland

Corresponding author:
Marika Kosiel-Pająk,
Women In International Security,
ul. Oleandrów 6, 00–629
Warsaw, Poland;
ORCID: 0000-0002-1187-5895;
marikakosiel@gmail.com

Cite this article as: M. Kosiel-Pająk, "UK Border Digitalisation – a Commentary on the Current State of Affairs," ACIG, vol. 1, no. 1, pp. 285–291, 2022.
DOI: 10.5604/01.3001.0016.1052

1. Background

A multifaceted and wholescale immigration reform was brought about by Brexit. It includes a plan to digitalise the immigration procedures by 2025, support eGates and eventually enforce a contactless mode of arrival (as for the majority of travellers). The digitisation of border control is part of a broader horizontal strategy drafted in the policy paper entitled 'Plan for Britain', published on 1 March 2017 [1]. The way the British government envisages the post-Brexit situation makes clear its desire to retain Britain's global role:

"The UK has a proud history of digital innovation: from the earliest days of computing to the development of the World Wide Web, the UK has been a cradle for inventions which have changed the world. And from Ada Lovelace – widely recognised as the first computer programmer – to the pioneers of today's revolution in artificial intelligence, the UK has always been at the forefront of invention."

The main aim of general digitalisation as a mainstream policy is presented as not only a way to a richer nation but also a more egalitarian one. Whereas, in fact, research and polls reveal that Britain is still a very class-ridden society with titled aristocrats, inherited wealth and privileges [2]. On the one hand mobility is thus an important drive towards greater social equality, which can be enhanced with proper migration management. On the other hand, the current system is definitely not fulfilling expectations, since it requires advanced digital literacy and visa waiting times are worsening as the backlog of applications rises.

The first stage after leaving behind the EU freedom of movement was registration of EU citizens that were settled in the UK prior to Brexit, and at the same time the introduction of the points-based migration system for any newcomers, which was intended to be: fair, firm and skills-driven. The pilot of the first fully digital migration route was already in place as of 2021 – knows as the EU Settlement Scheme. The system is currently functioning for almost 7 million applicants and has been slightly adapted on the way due to much higher demand than expected and complications arising from COVID-19.

The migration databases' interconnection with criminal records and the speeding up of procedures are also a must, with the new Electronic Travel Authorisation project planned for launch in 2023. The Home Office assumes that the system will become fully operational by 2024 and that an estimated 30 million non-visa visitors will apply electronically every year to be able to cross the border.

2. Serious delays in visa and passport processing

It is worth noting that currently the Home Office is dealing with certain procedures in premises in the UK, while some services, including biometric data collection, are already outsourced nationally to UK Visa and Citizenship Application Services (Sopra Steria) and worldwide Visa Centres (TLSContact, VFSGlobal).

According to the last update in August 2022 [3] – most of the entry schemes that are to be approached from outside the UK take about 8 weeks until completion, from the initial appointment in a visa centre to the issuance of the final decision. The goal is to have them finished in no more than 3 weeks. Some of the migration routes are even more time-consuming, with an average processing time of 12 weeks for a Hong Kong British National (Overseas) visa or family reunion. The longest procedures take an average of 24 weeks, as in the case of applying to settle in the UK as the spouse, partner or family member of someone who has British citizenship or is settled in the UK.

An applicant generally receives a decision on their UK visa application within 8 weeks (when applying from inside the UK), with the exception of the EU Settlement Scheme, which takes an average of 6 months. Special routes for Ukrainian citizens were criticised for taking months, especially in comparison with the EU's instant activation of the Temporary Protection Directive (Directive 2001/55/EC). Similarly, the UK Passport Office is dealing with a huge backlog when it comes to issuing new documents. The previous timeframe of 3 weeks (before the pandemic) has changed to 4–6 weeks or longer. The majority of passports are issued within 10 weeks [4].

The above examples clearly show that the goals of shrinking visa backlogs are quite ambitious and may not be answered solely by digitalisation. In fact, the expected waiting times are becoming longer and longer, rather than actually decreasing. Moreover, in practice it is difficult to receive any feedback on anticipated processing times in individual cases. This is especially true when there are no deadlines specified for various standard procedures on the part of British institutions. This process has a serious impact on the applicants involved, as their national passports are either retained or else there is no means of entering or exiting the UK throughout the whole immigration procedure (which may result in months or even years of uncertainty). Thus, the possibility of fast-tracking administrative procedures or judicial reviews (i.e. to challenge the legality of decisions or inaction in processing claims) is a proposal de lege ferenda (with a view to future law) for all migration routes.

3. Digitalisation process in policyplanning and implementation

The UK Government is presenting its policy rollout as an unbridled success, at least on paper. Even the former difficulties in execution due to COVID-19 or the need to add new routes are labelled as an uninterrupted string of successes. The Secretary of State for the Home Office, Priti Patel, published in July 2022 a New Plan for Immigration, in which she specified that: "Since this Government ended freedom of movement, we have implemented a global points-based system which has gone from strength to strength, with visa application numbers above those of pre-pandemic levels."

Moreover, the document further states boldly that "the changes we have already made enabled us to quickly adapt to meet the needs of those impacted by major international events such as the Covid-19 pandemic and the ongoing Russian invasion of Ukraine. We want to go

further, with customers experiencing a slick digital interface akin to their daily interactions with businesses or services" [5]. However, in spite of the pressure to deliver, both the parliamentary opposition and non-governmental organisations see many flaws that contradict this rosy view of affairs. A selection of examples of the most challenging issues in the ongoing digitisation process is provided below.

The commencement of the reform was very unfortunate, with critical failures in data security. Not long after launching the 2025 UK Border Strategy in December 2020 [6], due to a coding error within the Home Office, more than 213,000 fingerprint, DNA and arrest history records were erased from the Police National Computer. The inadvertent destruction of biometric evidence and the shutdown of the visa system software for 2 days caused acute embarrassment to the government and created a serious crisis for the system, if not a threat to public safety itself. Home Secretary Priti Patel attributed the critical situation in both justice and migration systems to a software bug [7].

4. Piloting the Border Crossing BX – EUSS as a digital system

The testing ground for the advanced digitalisation of procedures was primarily the registration system for EU residents – as a consequence of Brexit and based on Part Two of the UK-EU Withdrawal Agreement (European Union Settlement Scheme). The Agreement safeguards the post-Brexit residency rights of EU citizens and their families living in the UK (and *vice versa* for British citizens living in the EU countries) before the end of the transition period (set as 31 December 2020).

The most recent statistics show 6.7 million applications were received as of 30 June 2022, 91% of which were accounted for by European Economic Area (EEA) nationals, with the highest number of applications from Romanian (1,290,850), Polish (1,159,740) and Italian (594,390) nationals. The applicants could use their smartphones with NFC (Near-Field Communication) technology and the EU Exit: ID Document Check app to complete the identity verification stage of the procedure. The mobile app was used to check that an identity document is genuine and to confirm the bearer [8]. The EUSS was a step towards moving away from physical documents that confirm immigration status (stamps in passports and Biometric Residence Permits), and replacing such documents with an online immigration status verification system. The result of this procedure is digital – there is no physical proof of the right to a settled or pre-settled status. Therefore, a successful applicant demonstrates their status with share codes. The codes are generated online and may be shared with employers (right to work), landlords (right to rent), authorities (right to claim benefits) and law enforcement (right to remain), alongside the date of birth of the immigrant in question. The share code should be valid for 90 days. This service is designed for those who have UK Visas and Immigration account. By having a UK Visas and Immigration online account and presenting an EU national ID or passport at the border, a right of entry can be proved. An applicant during the procedure should have the same capabilities and rights

as the person already granted an official status, and they should be able to prove their rights with an application certificate in PDF form. In theory, it should be a smooth and efficient and user-friendly process, a win-win situation for both administration and applicants.

However, the digital status has many downsides and its users have made numerous complaints about it. The applicants and their concerns are not placed at the forefront of policy [9]. Furthermore, the accessibility of information and transparency of the procedures are not adequately secured. Not only it is difficult for some people (disabled, vulnerable, elderly, technologically excluded or disadvantaged) to rely on virtual rights without any physical backup document [10]. It also creates many stressful circumstances, especially at the airport with carriers having unclear liabilities for passengers that are unable to prove their rights with a traditional document . Not to mention the risk of being exploited by the so-called advice sharks that set up an account, but then make the applicants dependent on their services and unable to regain control over their own accounts and data [11].

5. Next step – Electronic Travel Authorisation,improvement of eGates and contactless border

The introduction of the ETA (Electronic Travel Authorisation), a pre-travel permission system (similar to the American ESTA and European ETIAS), is scheduled to start from 2023. It involves several stages of implementation. Before travelling to the UK without an advance visa or a confirmed immigration status, a non-visa national visitor will have to apply electronically for a special e-permission. The ETA should be automated and basic immigration and criminal records checks performed prior to issuing approval (with complex and adverse decisions being made by trained caseworkers). Apart from securing the border and preventing the arrival of someone that would not obtain entry clearance, it should also give the authorities the tools for thorough and precise tracking of entries and exits. Currently, the data on visas and border crossings are fragmented across disparate systems. The Initial Status Analysis system combines data from several administrative sources. The inclusion of data on EU citizens (with their different statuses) is to be gradually incorporated [12]. Moreover, the pandemic took its toll on the accuracy of data. The travel and tourism statistics are generally published by The Office for National Statistics and are based on the International Passenger Survey. The ETA scheme is due to operate fully by the end of 2024 [13].

When it comes to eGates, there are currently over 270 automated frictionless self-service barriers at 15 air and rail ports. These devices are supervised by the Border Guard and use data stored in a chip of biometric passports, along with a photograph or fingerprint taken at the time of entry, to verify the passport holder's identity. Although the current minimum age to use them is 12 (with an accompanying adult), the plan is to ease travel difficulties by reducing it to 10 years of age [14], which would enable further accelerated passenger flow management. The goal of increasing automation at the border is to eventually test a fully contactless travel experience, with the ability to undergo automated border screening [5].

However, specific solutions for contactless border management have not yet been presented by the Home Office. There are only Digital, Data and Technology principles available [15].

6. Conclusions

The full digitalisation of border control by 2025 is undoubtedly a truly ambitious plan. Though a digital end-to-end journey is the goal of the Home Office, it is nevertheless emphasised that this ought to be achieved without compromising on security, which has always been a top priority. The recent political and economic situation does not provide a seamless framework for such a sweeping reform. The changes in government and serious internal issues make it harder to implement far-reaching changes that require time and stability.

If the Home Office succeeds, it would undeniably put the UK at the forefront of border management. However, the implementation has been far from smooth to date. Technology can certainly help the Home Office staff to be more productive, especially taking into account their current backlogs. Although it provides data and many instruments to adjust migration policy on a fairly timely basis, its flaws are apparent given the delays and technical problems the Home Office is struggling to deal with.

Britain is not aspiring to welcome migrants with open arms, neither those arriving after lengthy and costly visa procedures nor forced migration. Digitalisation has the potential to dehumanise visa and immigration procedures even further. Today's progressively hard-line border policy, in terms of a fortress-like approach to every migrant as a potential over-stayer and exploiter of public funds, is arguably counterproductive. With the British economy in trouble and suffering from labour market shortages, the UK may find it increasingly difficult to attract both global talent and the low-paid workers it needs. A traveller-centred approach, with legal safeguards for the individuals concerned, should not only be an expedient catchphrase but a genuine reform principle underlying the UK's digital border and migration policy.

References

[1] HM Government, *Our Plan for Britain*. [Online]. Available: https://www.gov.uk/government/publications/uk-digital-strategy/executive-summary. [Accessed: Sep. 7, 2022].

[2] D. Robson. (2016). *How important is social class in Britain today?*. [Online]. Available: https://www.bbc.com/future/article/20160406-how-much-does-social-class-matter-in-britain-today. [Accessed: Sep. 7, 2022].

[3] HM Government, *Visa waiting times.* [Online]. Available: https://www.gov.uk/government/collections/visa-decision-waiting-times. [Accessed: Sep. 7, 2022].

[4] The UK Parliament, *HM Passport Office Backlogs*. [Online]. Available: https://hansard.parliament.uk/commons/2022-05-12/debates/6E89196F-813E-43FE-A500-E88CBAA9ABD7/HMPassportOfficeBacklogs#contribution-132E8E06-6BD4-430B-A5BE-6D21D753B848. [Accessed: Sep. 7, 2022].

[5] HM Government, *New Plan for Immigration: Legal Migration and Border Control*. [Online].

Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1092488/CCS204_CCS0722436296-001_Plan_for_Immigration_E-Laying.pdf. [Accessed: Sep. 7, 2022].

[6] HM Government, *2025 UK Border Strategy*. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/945380/2025_UK_Border_Strategy.pdf. [Accessed: Sep. 7, 2022].

[7] F. Hamilton. (2021). *Home Office urged to explain 150,000 arrest records wiped in tech blunder*. [Online]. Available: https://www.thetimes.co.uk/article/150-000-arrest-records-wiped-in-tech-blunder-krhlf302h. [Accessed: Sep. 7, 2022].

[8] HM Government. (2019). *2025 Using the 'EU Exit: ID Document Check' app.* [Online]. Available: https://www.gov.uk/guidance/using-the-eu-exit-Ad-document-check-app. [Accessed: Sep. 7, 2022].

[9] D.-O. Vicol. (2022). *EUSS digital-only status remains an issue for 2 in 5 EU citizens*. [Online]. Available: https://www.workrightscentre.org/news/euss-digital-only-status-remains-an-issue-for-2-in-5-eu-citizens. [Accessed: Sep. 7, 2022].

[10] J. Tomlinson, A. Welsh. (2020). *Digital Immigration Status; A Monitoring Framework*. [Online]. Available: https://publiclawproject.org.uk/content/uploads/2020/10/PLP-Report-Digital-Immigration-Status.pdf. [Accessed: Sep. 7, 2022].

[11] C. Barnard. (2019). *Unsettled Status*. [Online]. Available: https://ukandeu.ac.uk/unsettled-status/. [Accessed: Sep. 7, 2022].

[12] HM Government. (2022). "Developments in Exit Checks". [Online]. Available: https://www.gov.uk/government/statistics/developments-in-exit-checks/developments-in-exit-checks. [Accessed: Sep. 7, 2022].

[13] HM Government. (2022). *Policy paper. Nationality and Borders Bill: Electronic Travel Authorisation factsheet.* [Online]. Available: https://www.gov.uk/government/publications/nationality-and-borders-bill-electronic-travel-authorisation-factsheet/nationality-and-borders-bill-electronic-travel-authorisation-factsheet. [Accessed: Sep. 7, 2022].

[14] HM Government. (2021). *Guide to faster travel through the UK border.* [Online]. Available: https://www.gov.uk/government/publications/coming-to-the-uk/faster-travel-through-the-uk-border. [Accessed: Sep. 7, 2022].

[15] S. Bourne. (2021). *Home Office Digital, Data and Technology Strategy 2024.* [Online]. Available: https://www.gov.uk/government/publications/home-office-digital-data-and-technology-strategy-2024/home-office-digital-data-and-technology-strategy-2024. [Accessed: Sep. 7, 2022].

Editorial Office

Applied Cybersecurity & Internet Governance

Kolska Street 12
01–045 Warsaw, Poland
E-mail: contact@acigjournal.pl

https://acigjournal.com
https://acigjournal.pl

Publisher

NASK – National Research Institute

Kolska Street 12
01–045 Warsaw, Poland

www.nask.pl

Editorial Office

Applied Cybersecurity & Internet Governance

Kolska Street 12
01–045 Warsaw, Poland
www.acigjournal.com

Publisher

NASK – National Research Institute

Kolska Street 12
01–045 Warsaw, Poland
www.nask.pl