Mitigating Key Cyber Threats in 2023

Benjamin Lim
University of Edinburgh
s2599925@ed.ac.uk

December 10, 2023

1. Introduction

In 1996, John Perry Barlow famously declared cyberspace to be the "new home of Mind", one that is independent from the physical world [1]. Two decades have passed, the cyberspace natives Barlow spoke about have grown up and amalgamated cyberspace into the physical world. We use the Internet to check the weather, book a ride to get to our destination, navigate through the road network, order food at the restaurant and finally to pay for our meal. *A fortiori*, it is thus unsurprising that ENISA, in their 2023 Threat Landscape Report, observed an increase in cybersecurity attacks and their resulting consequences [2]. In this essay, I will explore how AI-enabled social engineering and Cyber Warfare between nation states have dominated the headlines in 2023 and dive into the various legal and technological means to mitigate these threats.

2. Introduction to Generative Artificial Intelligence (AI)

ChatGPT took the world by storm in 2022 with its ability to engage in human-like conversations. Academics were equally excited as it is slated to be one of the first AI algorithms to pass the Turing test since it was coined 70 years ago. To pass the Turing test, an AI's response would need to be virtually indistinguishable from a human's response. ChatGPT is but one of many Generative AI algorithms that have experienced a recent breakthrough. Other promising algorithms include Stable Diffusion, which can generate brand new images from a text description, as well as VALL-E, which can synthesize speech from a three second voice recording [3]. In the past, we could edit images with software or piece together short recordings of a person's voice. However, such techniques were extremely time consuming, required specialized skills and were likely to contain minor flaws which might reveal that the media has been doctored. In Lenihan v Shankar, a paternity test submitted to the court was discovered to be falsified as the footnotes had been accidentally overwritten when the editing was done [4]. Future Shankars' will be able to harness the power of Generative AI to craft flawless images that will pose a challenge to detecting *crimen falsi*. With so much potential for misuse, I would like to focus on the misuse of Generative AI for social engineering in the next section.

[1] John Perry Barlow, 'A Declaration of the Independence of Cyberspace' (*Electronic Frontier Foundation*, 8 February 1996) <https://www.eff.org/cyberspace-independence> accessed 30 October 2023

[2] European Union Agency for Cybersecurity, *ENISA Threat Landscape 2023* (2023) <https://www.enisa.europa.eu/topics/cyber-threats/threats-and-trends>, pp. 4

[3] Chengyi Wang and others, 'Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers' [2023] <https://doi.org/10.48550/arXiv.2301.02111>, pp. 1

[4] *Lenihan v Shankar* [2021] Ontario Superior Court of Justice 330, para. 202

3. AI generated phishing emails

We have all received phishing emails where the cybercriminal is trying to convince the victim to part with their hard-earned money through a variety of lures such as paying a small amount of tax to receive a supposed huge inheritance. One telltale sign is the presence of spelling and grammatical mistakes in the email. There are two schools of thought here. The first group believes the mistakes are due to poor linguistic ability of the cybercriminals. Generative AI algorithms can assist criminals to draft convincing emails with flawless language. The second group believes the mistakes are planted on purpose. Given the lucrative nature of the criminal operation, it is trivial to hire a fluent speaker to proofread the emails if they wanted to. *A priori*, the mistakes are planted to filter out educated victims so that only uneducated gullible victims respond to the email. Educated victims would discover the ploy after exchanging a few emails which would mean wasted resources responding to them. It would be more productive to target only uneducated gullible victims [5]. Generative AI algorithms can also assist this group of cybercriminals by automating email responses and allowing these criminals to cast a wider net. Given that Generative AI is on the verge of passing the Turing test, I believe it is possible for AI to craft relevant responses to the victim's questions without human intervention.

Generative AI can also craft new phishing scenarios or paraphrase existing ones in an attempt to evade widespread detection. In a corporate environment, when emails are reported and confirmed as phishing, security analysts are trained to search for and purge all emails from the same sender and with the same contents, so as to prevent others from falling victim [6]. Cybercriminals are already sending the same phishing emails from multiple different email accounts to avoid detection, hence filtering solely by sender is not sufficient. If generative AI is used to paraphrase the contents for every email sent out, it would pose a great challenge to existing email protection solutions as there is currently no good indicator to find all the emails sent out as part of the same campaign.

4. Addressing AI generated phishing emails

There exist both legal and technological measures to address the new risks posed by AI generated phishing emails. OpenAI has self-regulated and put in place "ethical guardrails" that prevent it from giving responses which are deemed unethical, such as how to commit murder or draft phishing emails [7]. Such guardrails are important because Generative AI algorithms are a dual use technology that can be harnessed for both *bona fide* and *mala in se* uses. Dual use technology exists in other fields as well. Certain controlled drugs like Morphine have positive effects but can be abused as well. Abuse of such drugs can be countered by requiring prescriptions before they can be dispensed. However, in the case of Generative AI algorithms, such countermeasures are not feasible. It is simply not possible for someone to vet every request, hence guardrails must be built into the algorithm. The UK's AI White Paper has also called out the specific security risks from hacker's use of Generative AI [8] and proposed that individual

[5] Daniel Wyatt and Christopher Whitehouse, 'What to know about AI fraudsters before facing disputes' (*Law360*, 31 August 2023) <https://plus.lexis.com/api/permalink/c1d849b7-8628-4ac1-b988-53dbaec9371f/?context=1001073> accessed 1 November 2023

[6] Microsoft, 'Remediate malicious email delivered in Office 365' (*Microsoft*, 19 June 2023) <https://learn.microsoft.com/en-us/microsoft-365/security/office-365-security/remediate-malicious-email-delivered-office-365?view=o365-worldwide> accessed 31 October 2023

[7] Wyatt and Whitehouse (n 5)

[8] Department for Science, Innovation & Technology, *A pro-innovation approach to AI regulation* (2023) <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper>, Section 1.2

sectoral regulators implement it into their existing regulatory frameworks. While some ethical companies can be relied on to self-regulate, government regulation is crucial to ensure minimum standards across the board and retain public trust in AI technologies. These are all valiant efforts to prevent the misuse of AI for phishing. Nonetheless, hackers have found technological workarounds. Since ChatGPT allows the user to instruct it to generate certain content, hackers have found ways to instruct it to ignore the guardrails and generate content that would have been proscribed [9]. It remains to be seen if ChatGPT is able to set those tenets in stone such that hackers can never reprogram it to ignore those guardrails.

Aside from preventing the AI from generating phishing emails, measures to filter out and block AI generated phishing emails would also mitigate the issue. Since the email subject can be repeatedly paraphrased by AI, a simple matching solution will not work. Security engineers could use AI algorithms like the Naive Bayes classifier to generate a list of similar words to detect paraphrased emails [10]. Such a method will allow analysts to detect similar emails in the same phishing campaign that have been paraphrased. Using AI to defend against AI is simply poetic justice. While *prima facie* effective, such methods might also introduce false positives. The word cloud below shows similar words that might be used in a Nigerian Prince scam. However, these same words are also commonly used in a banking environment, hence legitimate emails may also get mistakenly categorized as malicious. It will require fine-tuning and further development to achieve the same level of phishing email detection as before. Cybersecurity is a dynamic field where battle lines are frequently redrawn. There may come a day where we can no longer reliably purge AI paraphrased email and we have to instead rely more heavily on educating users to identify phishing emails.

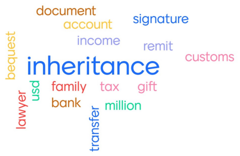

Figure 1: Similar words displayed in a Word Cloud. [12]

Using AI generated Deepfakes for phishing

The proverb "seeing is believing" can be traced back to the 18th century. Unfortunately, it may soon meet its demise given the recent advances in Generative AI. Algorithms like Stable Diffusion allow anyone to generate images and scenes that never happened in reality from just a text prompt. Such images are known as deepfakes due to the high quality and convincibility. Cybercriminals can use deepfakes of famous individuals to promote their products and boost their credibility or build a convincing website which is able to fool victims into providing their personal information. Deepfakes can also be used to generate incredulous images to be used as bait, enticing victims to click on the link to find out more about shocking news.

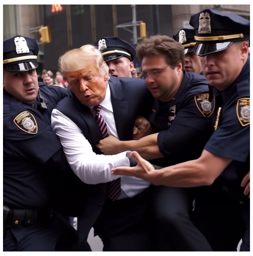

Figure 2: Deepfake of Trump being arrested. [14]

[9] Michael King, 'Meet DAN — The 'JAILBREAK' Version of ChatGPT and How to Use it — AI Unchained and Unfiltered' (*Medium*, 5 February 2023) <https://medium.com/neonforge/meet-dan-the-jailbreak-version-of-chatgpt-and-how-to-use-it-ai-unchained-and-unfiltered-f91bfa679024> accessed 1 November 2023

[10] Sheila Farach Diba and Jaka Nugraha, 'Implementation of Naive Bayes Classification Method for Sentiment Analysis on Community Opinion to Indonesian Criminal Code Draft' (2019) 474 Advances in Social Science, Education and Humanities Research <https://www.atlantis-press.com/article/125944919.pdf>

[12] Generated using Mentimeter

Apart from images, deepfakes of a person's voice can also be generated using algorithms like VALL-E. Voice deepfakes are used in Vishing (Voice Phishing) attempts. These are normally more targeted and criminals would spend time understanding the target company's business and reporting structure. Once an opportunity arises, such as management being away on vacation, the criminals would strike and call the employee using a generated recording of the manager's voice instructing the employees to perform an urgent fund transfer immediately. Attempts to reach out to the manager might go unanswered since he or she is on vacation and the employee may make that transfer under duress.

5. Addressing use of AI generated Deepfakes for phishing

Apart from "ethical guardrails" as well as the UK's AI white paper which would also apply to deepfakes, other measures include the DEEPFAKES Accountability Bill in the US, which would require deepfakes to be *inter alia* watermarked and accompanied by disclosures [15]. Critics argued that the bill was overly broad and covered videos created via any "technical means", which included regular video editing techniques like slowing down a video [16]. Since there were no reliable methods to detect deepfakes and to differentiate between *bona fide* and *mala in se* uses of deepfakes, there were concerns whether the bill could be enforced [17]. The bill was floated during election season due to the potential of deepfakes disrupting the election and was subsequently shelved after the election passed due to dissipating interest.

Existing legal instruments can also be used to combat deepfakes and phishing. To combat phishing, the law can target financial intermediaries to prevent the transfer of funds to fraudsters. In Philipp v Barclays, law enforcement notified the bank that the victim's account had been compromised which led to the bank freezing the account and preventing further losses [18]. Money mules have also been prosecuted for their role in facilitating the transfer of funds from victims to criminals. In R v Krishnasamy, the defendant was convicted under the Proceeds of Crime Act [19] for his role as a "mule herder", transferring funds out of "mule" accounts for a commission [20]. The funds were illegally obtained through payment diversion fraud [21], where fraudsters impersonate someone's identity through methods such as vishing and convince the victim to divert the payment into the mule account. Intermediaries such as hosting providers can also be targeted to takedown websites so as to prevent public from being misled by the deepfakes into providing their personal information. A multinational law enforcement effort led to the *in rem* seizure of USD 8.6 million and the takedown of infrastructure supporting Qakbot malware which infected over 700,000 victims through methods

[14] @EliotHiggins (https://twitter.com/EliotHiggins/status/1637928223848767492/photo/2) accessed 17 November 2023

[15] DEEPFAKES Accountability Bill (2019–20), section 2

[16] Zachary Schapiro, 'Deep fakes accountability act: Overbroad and ineffective' [2020] Boston College Intellectual Property and Technology Forum <https://lira.bc.edu/work/ns/73aa6d56-3d4b-4176-bf20-446281904b04>, pp. 7

[17] Schapiro (n 16), pp. 11

[18] *Philipp v Barclays Bank UK plc* (2023) 3 WLR 284, para. 13

[19] Proceeds of Crime Act 2002

[20] *R v Krishnasamy* [2021] EWCA Crim 232, para. 15

[21] *R v Krishnasamy* (n 20), para. 11

such as phishing [22]. Lastly, the Defamation Act provides recourse for individuals who can obtain an injunction to prevent further usage of deepfakes in their image [23]. That said, the injunction is unlikely to deter criminals who are already using the deepfakes for illegal purposes.

Technological measures are also crucial to mitigate the negative impact of deepfakes. The signaling system No. 7 (SS7) protocol was developed in the 1970s to enable phone calls and SMS to be exchanged between a private network of trusted mobile operators. As the world gradually became more interconnected, the number of operators that were part of the network grew. Soon, criminals gained access to the network by hacking insecure operators [24], and were able to make and receive phone calls or SMS that appeared to originate from any phone number [25]. Deepfakes, together with SS7 vulnerabilities, have allowed criminals to make convincing vishing calls. Newer protocols such as Voice over Internet Protocol (VoIP) or Signal are secure by design. They require strong authentication and thus prevent identity spoofing. Vishing calls made over secure protocols would appear to originate from an unknown person and thus not be as convincing as traditional calls. By moving to secure communication applications and through continual user education, we will be able to mitigate the impact of vishing attempts.

6. Cyberspace as a new battleground

The world has been fraught with geopolitical tension and conflict recently. With the increasing interconnectedness of the digital world to the physical world, it is no surprise that such conflict has also found its way into cyberspace. With the backing of a nation state, such attacks are unprecedented in term of variety and complexity, resulting in collateral damage to civilian infrastructure which stand little chance [26]. Governments may have to step in to defend against such sophisticated cyber attacks, further blurring the line between public and private responsibility. In this section, I will explore how Remote Access Trojans (RATs) and distributed denial of service (DDoS) activity have turned cyberspace into the next battleground.

7. RATs in Cyber Warfare

Nitzberg has defined Information Warfare as the "use (and abuse) of computers [...] to undermine the computing resource of an adversary" [27]. With the Russian war against Ukraine, the Russian-based Turla APT group was observed to have stepped up their attacks on Ukrainian defense targets [28]. Turla uses sophisticated custom developed RATs such as "snake" to steal confidential documents from infected comput-

[22] US Department of Justice, 'Qakbot Malware Infected More Than 700,000 Victim Computers, Facilitated Ransomware Deployments, and Caused Hundreds of Millions of Dollars in Damage' (*US Department of Justice*, 29 August 2023) <https://www.justice.gov/usao-cdca/pr/qakbot-malware-disrupted-international-cyber-takedown> accessed 3 November 2023

[23] Defamation Act 2013

[24] Hassan Mourad, 'The Fall of SS7 – How Can the Critical Security Controls Help?' [2015] Global Information Assurance Certification Paper <https://www.giac.org/paper/gccc/276/fall-ss7-critical-security-controls-help/119644>, pp. 6

[25] Mourad (n 24), pp. 6-11

[26] Scott J Shackelford, *Managing Cyber Attacks in International Law, Business, and Relations* (Cambridge University Press 2014), pp. 172-173

[27] Sam Nitzberg, 'Conflict and the Computer: Information Warfare and Related Ethical Issues' [1998] Proceedings of the 21st National Information Systems Security Conference <https://csrc.nist.gov/files/pubs/conference/1998/10/08/proceedings-of-the-21st-nissc-1998/final/docs/paperd7.pdf>, pp. 55

[28] Daryna Antoniuk, 'Russia's Turla hackers target Ukraine's defense with spyware' (*Recorded Future*, 19 July 2023) <https://therecord.media/turla-hackers-targeting-ukraine-defense> accessed 11 November 2023

ers. Unlike financial theft or theft of physical goods, making copies of confidential documents does not affect the original. Hence such malware can remain hidden for years, continuing to pilfer new documents regularly. Such documents are then analyzed for intelligence about military activity and capabilities which are weaponized in subsequent attacks. The potential impact can be substantial as the intelligence can be used to influence decision making that might result in loss of life during actual combat. Furthermore, as I will explore later, a leak of these sophisticated tools into the public domain may cause a spillover effect when criminals leverage these tools to conduct attacks on the general public.

8. Addressing RATs in Cyber Warfare

Such egregious malware can be addressed through a mixture of legal and technological measures such as search warrants and custom developed tools. To tackle Turla, the FBI obtained a search warrant to deploy a custom created tool which issued commands to "Snake" to overwrite itself [29]. Using purely legal instruments would not have sufficed since the other party is a Nation State alleged to be behind the attack. Even if treaties exist, "enforcement [would be] a problem" [30]. Purely technological measures would also not suffice as such actions would constitute "unauthorized access" to the victim's computer which would be illegal in most countries. Hence, tackling such threats requires the use of legal instruments like court orders to authorize the deployment of technological tools. Since Turla targeted computers from at least 50 countries, transnational cooperation between friendly nation states would be required to obtain *lex loci* authorization to deploy the tool in their countries [31].

While such actions might be effective in addressing RAT activity, it creates increasing polarization in the global context, forcing countries to take sides to obtain protection from countries with the technological means to develop sophisticated tools and detect such advanced threats. The United States have been caught with their hands in the cookie jar as well. Custom hacking tools, alleged to have been developed by the National Security Agency (NSA), were leaked by a group called Shadow Brokers [32]. The increased militarization of cyberspace with countries all trying to build their own arsenal of custom tools further contributes to the increased polarization.

A leak of these sophisticated state funded cyberweapons into the public domain will cause mayhem due to the imbalance in the cybersecurity capabilities of private sector compared to the deep pockets of defence establishment. It is akin to giving criminals access to missiles which can be freely duplicated, for which the general public has no defence against. The aforementioned leak by Shadow Brokers included an exploit known as "EternalBlue" which could target practically any computer running Windows. This exploit was re-purposed by an allegedly North Korean criminal group to deploy WannaCry ransomware worldwide, causing an estimated 4 billion USD in recovery costs [33]. The exploit was so potent that Microsoft took a "highly unusual" step of issuing patches even for operating sys-

[29] USAttorney's Office, Eastern District of New York, 'Justice Department Announces Court-Authorized Disruption of the Snake Malware Network Controlled by Russia's Federal Security Service' (*US Department of Justice*, 19 July 2023) <https://www.justice.gov/usao-edny/pr/justice-department-announces-court-authorized-disruption-snake-malware-network> accessed 11 November 2023

[30] Shackelford (n 26), pp. 91

[31] USAttorney's Office, Eastern District of New York (n 29)

[32] Matt Burgress, 'Hacking the hackers: everything you need to know about Shadow Brokers' attack on the NSA' (*WIRED*, 10 April 2017) <https://www.wired.co.uk/article/nsa-hacking-tools-stolen-hackers> accessed 11 November 2023

[33] Jennifer Gregory, 'What Has Changed Since the 2017 WannaCry Ransomware Attack?' (*IBM*) <https://securityintelligence.com/articles/what-has-changed-since-wannacry-ransomware-attack/> accessed 11 November 2023

tems which were no longer officially supported [34].

Transnational legal instruments will help encourage "mutual disarmament" by harmonizing vulnerability disclosure across various jurisdictions. In the aftermath of WannaCry, the GCHQ made the courageous step of disclosing "BlueKeep", a slightly less potent "god mode" exploit [35]. A number of other countries are also abiding by a Vulnerabilities Equities Process (VEP) to guide decision making on disclosure of vulnerabilities. The VEP is a framework to help intelligence agencies weigh the benefits of keeping vulnerabilities secret so they can be used for intelligence collection purposes against the possible impacts that might occur if the vulnerability is leaked or if another country manages to independently discover the same vulnerability and use it against them [36]. Hopefully, VEPs can be established as a cyber norm and harmonized across more jurisdictions.

9. Cyber Warfare DDoS activity

Apart from RAT activity, there was also an uptick in DDoS activity observed on Ukraine's banking sector that has been attributed to Russian sources [37]. DDoS attacks make use of botnets as well as certain vulnerable services to generate a huge volume of traffic to overwhelm the target such that legitimate traffic is unable to get through. Botnets are a collection of computers that have been previously compromised and thus can be commandeered by the attackers to flood the target with meaningless traffic. Certain services such as DNS may be vulnerable to a form of an attack known as an amplification attack [38] [39]. An amplification attack requires the attacker to spoof his IP address as the target and send a small amount of traffic to the vulnerable service which would generate a huge response DDoSing the actual target. As a result of the DDoS activity, there was intermittent disruption and Ukranians were unable to use the ATMs and their bank cards temporarily [40].

10. Addressing Cyber Warfare DDoS activity

The RAT activity previously discussed targeted the defense sector, which is probably *jus in bello* conduct, since it has the equivalent result "as a physical invasion by traditional military forces [41]. However, this DDoS activity targeted the financial sector, resulting in collateral damage to civilians who were temporarily unable to access critical services. There were a number of legal instruments that were supposed to prevent the

[34] Microsoft, 'Customer Guidance for WannaCrypt attacks' (*Microsoft*) <https://msrc.microsoft.com/blog/2017/05/customer-guidance-for-wannacrypt-attacks/> accessed 11 November 2023

[35] Nicholas Weaver, 'The GCHQ's Vulnerabilities Equities Process' (*Lawfare*) <https://www.lawfaremedia.org/article/gchqs-vulnerabilities-equities-process> accessed 11 November 2023

[36] The White House, *Vulnerabilities Equities Policy and Process for the United States Government* (2017) <https://trumpwhitehouse.archives.gov/sites/whitehouse.gov/files/images/External%20-%20Unclassified%20VEP%20Charter%20FINAL.PDF>

[37] National Cybersecurity Centre, 'UK assesses Russian involvement in cyber attacks on Ukraine' (*National Cybersecurity Centre*, 18 February 2022) <https://www.gov.uk/government/news/uk-assess-russian-involvement-in-cyber-attacks-on-ukraine> accessed 11 November 2023

[38] Cloudflare, 'DNS amplification attack' (*Cloudflare*) <https://www.cloudflare.com/en-gb/learning/ddos/dns-amplification-ddos-attack/> accessed 11 November 2023

[39] Internet Society, *Addressing the challenge of IP spoofing* (2015) <https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-AntiSpoofing-20150909-en-2.pdf>, pp. 3

[40] Jenna McLaughlin, 'Ukraine says government websites and banks were hit with denial of service attack' (*National Public Radio*, 15 February 2022) <https://www.npr.org/2022/02/15/1080876311/ukraine-hack-denial-of-service-attack-defense> accessed 11 November 2023

[41] Shackelford (n 26), pp. 287

occurrence of such activity. Firstly the UN cyber norms, of which Russia participated in the drafting, remarks that States must not use proxies to commit malicious ICT acts and ensure their territories are not used by other actors to commit such acts [42]. Secondly, the Organization for Security and Co-operation in Europe (OSCE), which includes both Russia and Ukraine, have agreed on a series of Confidence-Building measures, which include holding consultations to reduce probability of conflict and to protect critical ICT infrastructure [43]. Unfortunately, these supranational agreements failed to prevent the DDoS attack on Ukraine's banking sector.

Hope is not lost, legal instruments such as the upcoming EU Cybersecurity Certification scheme will provide assurance that certain products have been tested for vulnerabilities [44] and are *prima facie* secure. Consumers will be able to make a more informed decision to choose these certified products which would reduce the number of vulnerable devices on the Internet, thus reducing the size of botnets. Furthermore, certain countries such as Finland have mandated that Internet Service Providers (ISPs) put in place measures to prevent source IP spoofing [45] [46], hence preventing hackers from using their networks for amplification attacks. Apart from legislation, the Internet Society has spearheaded an initiative, Mutually Agreed Norms for Routing Security (MANRS), to encourage all ISPs to play their part in securing their networks so it cannot be used as a launchpad to conduct amplification attacks. [47]. Hence, with the reduction in botnet size as well as more ISPs implementing anti IP spoofing measures, the impact of future DDoS activity will be reduced.

Technological measures also feature in mitigating DDoS activity from a Nation State actor. ISPs can play a big role by blackholing malicious traffic at the Internet exchange point (IXP) before it enters the country's domestic network. The ISPs have a much larger bandwidth pipe and can more effectively handle higher volumes of traffic [48]. A reporting framework should be set up so key personnel in the critical infrastructure sectors can make contact with the ISPs to swiftly share information about ongoing DDoS attacks so that traffic patterns can be determined, and filters added to separate genuine and malicious data packets [49]. Traffic patterns may include certain packet sizes or headers that legitimate traffic is unlikely to use. Rate limiting can be another effective strategy if it proves impossible to distinguish malicious traffic from legitimate traffic. Malicious traffic would be throttled, and the service will not be overwhelmed. Since legitimate traffic is unlikely to be voluminous, it will not be too adversely affected by the rate limit.

While necessary, setting up filtering mechanisms at the IXP where Internet traffic first

[42] Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, UNGA (24 June 2013) UN Doc A/68/98 (2013) <https://digitallibrary.un.org/record/753055?ln=en>, pp. 8

[43] Decision No. 1202 OSCE Confidence-Building Measures To Reduce the Risks of Conflict Stemming From the Use of Information and Communication Technologies, Organization for Security and Co-operation in Europe (10 March 2016) PC.DEC/1202 <https://www.osce.org/files/f/documents/d/a/227281.pdf>

[44] European Union Agency for Cybersecurity, *Accreditation of ITSEFs for the EUCC scheme* (2013) <https://certification.enisa.europa.eu/documentation/eucc_state_of_the_art_accreditation-of-itsefs-v1-1.pdf>

[45] Määräys teletoiminnan tietoturvasta 2015, para. 11

[46] Internet Society (n 39), pp. 14

[47] Internet Society (n 39), pp. 14

[48] Cloudflare (n 38)

[49] Pardis Moslemzadeh Tehrani, *Cyberterrorism: The Legal and Enforcement Issues* (World Scientific 2017), pp. 224

[50] Barlow (n 1)

enters the country changes the nature of the Internet. Barlow once declared that "Cyberspace does not lie within your borders" [50]. Today, with cyberspace so tightly integrated with the physical world, traditional geographical borders have to be set up in cyberspace especially in times of conflict where the adversary is intent on using cyberspace as a battleground.

11. Conclusion

To conclude, we have critically analyzed two key cybersecurity threats in this essay, AI-enabled social engineering as well as Cyber Warfare. Generative AI can be used to craft emails and to create deepfakes for use in phishing. We have also looked at legal measures such as "ethical guardrails", the DEEPFAKES accountability bill as well as the use of intermediaries to mitigate the effect of AI-enabled social engineering. Technological measures such as using AI for defense, using newer secure by design protocols were also proposed as potential solutions.

On the Cyber Warfare front, we have explored how RATs and DDoS attacks are some of the techniques utilized by nation state threat actors today. Legal instruments such as the UN cyber norms and the OSCE Confidence-Building measures were critically evaluated. A hybrid legal-technological solution involving using court orders to authorize deployment of tools was found to be more effective. We also explored how criminals could use leaked cyberweapons to wreck widespread damage on the Internet. Lastly, we considered the role ISPs could play to mitigate DDoS activity from Nation State actors.

In the dynamic cybersecurity threat landscape today, new technologies like Generative AI could rapidly disrupt and render existing controls ineffective. Sudden geopolitical developments could also lead to a breakdown in cooperation between various governments and render existing legal instruments otiose. As cyberspace continues to integrate with the physical world, it is inevitable that it would eventually fragment and take on the same geopolitical boundaries present in the physical world. It is more crucial today than ever before to explore a combination of new legal and technological solutions to mitigate these new cybersecurity threats to minimize the disruption caused to the general public, who have become so reliant on the continued usage of technologies in cyberspace.

Bibliography

Cases

Lenihan v Shankar [2021] Ontario Superior Court of Justice 330.

Philipp v Barclays Bank UK plc (2023) 3 WLR 284.

R v Krishnasamy [2021] EWCA Crim 232.

Legislation

DEEPFAKES Accountability Bill (2019–20).

Defamation Act 2013.

Määräys teletoiminnan tietoturvasta 2015.

Proceeds of Crime Act 2002.

Reports

Department for Science, Innovation & Technology, *A pro-innovation approach to AI regulation* (2023) <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper>.

European Union Agency for Cybersecurity, *Accreditation of ITSEFs for the EUCC scheme* (2013) <https://certification.enisa.europa.eu/documentation/eucc_state_of_the_art_accreditation-of-itsefs-v1-1.pdf>.

— *ENISA Threat Landscape 2023* (2023) <https://www.enisa.europa.eu/topics/cyber-threats/threats-and-trends>.

Internet Society, *Addressing the challenge of IP spoofing* (2015) <https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-AntiSpoofing-20150909-en-2.pdf>.

The White House, *Vulnerabilities Equities Policy and Process for the United States Government* (2017) <https://trumpwhitehouse.archives.gov/sites/whitehouse.gov/files/images/External%20-%20Unclassified%20VEP%20Charter%20FINAL.PDF>.

Books

Shackelford SJ, *Managing Cyber Attacks in International Law, Business, and Relations* (Cambridge University Press 2014).

Tehrani PM, *Cyberterrorism: The Legal and Enforcement Issues* (World Scientific 2017).

Articles

Diba SF and Nugraha J, 'Implementation of Naive Bayes Classification Method for Sentiment Analysis on Community Opinion to Indonesian Criminal Code Draft' (2019) 474 Advances in Social Science, Education and Humanities Research <https://www.atlantis-press.com/article/125944919.pdf>.

Mourad H, 'The Fall of SS7 – How Can the Critical Security Controls Help?' [2015] Global Information Assurance Certification Paper <https://www.giac.org/paper/gccc/276/fall-ss7-critical-security-controls-help/119644>.

Nitzberg S, 'Conflict and the Computer: Information Warfare and Related Ethical Issues' [1998] Proceedings of the 21st National Information Systems Security Conference <https://csrc.nist.gov/files/pubs/conference/1998/10/08/proceedings-of-the-21st-nissc-1998/final/docs/paperd7.pdf>.

Schapiro Z, 'Deep fakes accountability act: Overbroad and ineffective' [2020] Boston College Intellectual Property and Technology Forum <https://lira.bc.edu/work/ns/73aa6d56-3d4b-4176-bf20-446281904b04>.

Wang C and others, 'Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers' [2023] <https://doi.org/10.48550/arXiv.2301.02111>.

Secondary Sources

Antoniuk D, 'Russia's Turla hackers target Ukraine's defense with spyware' (*Recorded Future*, 19 July 2023) <https://therecord.media/turla-hackers-targeting-ukraine-defense> accessed 11 November 2023.

Barlow JP, 'A Declaration of the Independence of Cyberspace' (*Electronic Frontier Foundation*, 8 February 1996) <https://www.eff.org/cyberspace-independence> accessed 30 October 2023.

Burgress M, 'Hacking the hackers: everything you need to know about Shadow Brokers' attack on the NSA' (*WIRED*, 10 April 2017) <https://www.wired.co.uk/article/nsa-hacking-tools-stolen-hackers> accessed 11 November 2023.

Cloudflare, 'DNS amplification attack' (*Cloudflare*) <https://www.cloudflare.com/en-gb/learning/ddos/dns-amplification-ddos-attack/> accessed 11 November 2023.

Gregory J, 'What Has Changed Since the 2017 WannaCry Ransomware Attack?' (*IBM*) <https://securityintelligence.com/articles/what-has-changed-since-wannacry-ransomware-attack/> accessed 11 November 2023.

King M, 'Meet DAN — The 'JAILBREAK' Version of ChatGPT and How to Use it — AI Unchained and Unfiltered' (*Medium*, 5 February 2023) <https://medium.com/neonforge/meet-dan-the-jailbreak-version-of-chatgpt-and-how-to-use-it-ai-unchained-and-unfiltered-f91bfa679024> accessed 1 November 2023.

McLaughlin J, 'Ukraine says government websites and banks were hit with denial of service attack' (*National Public Radio*, 15 February 2022) <https://www.npr.org/2022/02/15/1080876311/ukraine-hack-denial-of-service-attack-defense> accessed 11 November 2023.

Microsoft, 'Remediate malicious email delivered in Office 365' (*Microsoft*, 19 June 2023) <https://learn.microsoft.com/en-us/microsoft-365/security/office-365-security/remediate-malicious-email-delivered-office-365?view=o365-worldwide> accessed 31 October 2023.

— 'Customer Guidance for WannaCrypt attacks' (*Microsoft*) <https://msrc.microsoft.com/blog/2017/05/customer-guidance-for-wannacrypt-attacks/> accessed 11 November 2023.

National Cybersecurity Centre, 'UK assesses Russian involvement in cyber attacks on Ukraine' (*National Cybersecurity Centre*, 18 February 2022) <https://www.gov.uk/government/news/uk-assess-russian-involvement-in-cyber-attacks-on-ukraine> accessed 11 November 2023.

USAttorney's Office, Eastern District of New York, 'Justice Department Announces Court-Authorized Disruption of the Snake Malware Network Controlled by Russia's Federal Security Service' (*US Department of Justice*, 19 July 2023) <https://www.justice.gov/usao-edny/pr/justice-department-announces-court-authorized-disruption-snake-malware-network> accessed 11 November 2023.

US Department of Justice, 'Qakbot Malware Infected More Than 700,000 Victim Computers, Facilitated Ransomware Deployments, and Caused Hundreds of Millions of Dollars in Damage'

(*US Department of Justice*, 29 August 2023) <https://www.justice.gov/usao-cdca/pr/qakbot-malware-disrupted-international-cyber-takedown> accessed 3 November 2023.

Weaver N, 'The GCHQ's Vulnerabilities Equities Process' (*Lawfare*) <https://www.lawfaremedia.org/article/gchqs-vulnerabilities-equities-process> accessed 11 November 2023.

Wyatt D and Whitehouse C, 'What to know about AI fraudsters before facing disputes' (*Law360*, 31 August 2023) <https://plus.lexis.com/api/permalink/c1d849b7-8628-4ac1-b988-53dbaec9371f/?context=1001073> accessed 1 November 2023.