Module II

ARTIFICIAL INTELLIGENCE IN CYBERSECURITY

1. Introduction

According to many security analysts, security incidents reached the highest number ever recorded in 2019.[25] From phishing to ransomware, from dark web as a service economy to attacks on civil infrastructure, the cybersecurity landscape involved attacks that grew increasingly sophisticated during the year.[26] This upwards trend continued in 2020. The volumeof malware threats observed averaged 419 threats per minute, an increase of 44 threats per minute (12%) in the second quarter of 2020.[27] Cyber criminals managed to exploit the Covid- 19 pandemic and the growing online dependency of individuals and corporations, leveraging potential vulnerabilities of remote devices and bandwidth security. According to Interpol, 907,000 spam messages related to Covid-19 were detected between June and April 2020. Similarly, the 2020 Remote Workforce Cybersecurity Report showed that nearly two thirds of respondents saw an increase in breach attempts, with 34% of those surveyed having experienced a breach during the shift to telework.[28] Exploiting the potential for high impact and financial benefit, threat actors deployed themed phishing emails impersonating government and health authorities to steal personal data and deployed malware against criticalinfrastructure and healthcare institutions.[29]

In 2021 the drive for ubiquitous connectivity and digitalisation continues to support economicprogress but also, simultaneously and 'unavoidably', creates a fertile ground for the rise in scale and volume of cyberattacks. Increasing ransomware and diversified tactics, increasingly mobile cyber threats, ever more sophisticated phishing, cyber criminals and nation state attackers targeting the systems that run our day-to day-lives and malicious actors attacking the cloud forevery new low-hanging fruit. [30]

2. AI systems' support to cybersecurity

Against this backdrop, organisations have started using AI to help manage a growing range of cybersecurity risks, technical challenges, and resource constraints by enhancing their systems' robustness, resilience, and response. Police dogs provide a useful analogy to understand why companies are using AI to increase cybersecurity. Police officers use police dogs' specific abilities to hunt threats; likewise, AI systems work with security analysts to change the speed

---

[25] In the first quarter of 2019, businesses detected a 118% increase in ransomware attacks and discovered new ransomware families such as Anatova, Dharma and GandCrab, which use innovative techniques to target and infect enterprises, MacAfee (2019), "McAfee Labs Threats Report", August.

[26] M.Drolet (2020), "The Evolving Threat Landscape: Five Trends to Expect in 2020 and Beyond", Forbes Technology Council; Orange Business Service (2020), "2020 Security Landscape".

[27] MacAfee (2020), "McAfee Labs Threats Report", November.

[28] Fortinet (2020), Enterprises Must Adapt to Address Telework Security Challenges: 2020 Remote Workforce Cybersecurity Report", August.

[29] INTERPOL (2020), "INTERPOL report shows alarming rate of cyberattacks during COVID-19", August (www.interpol.int/en/News-and-Events/News/2020/INTERPOL-report-shows-alarming-rate-of-cyberattacks-during- COVID-19).

[30] Splunk (2019), "IT Security Predictions 2020"; ENISA (2020), "Emerging Trends – ENISA Threat Landscape", 20 October (www.enisa.europa.eu/publications/emerging-trends)

at which operations can be performed. In this regard, the relationship between AI systems andsecurity operators should be understood as a synergetic integration, in which the unique addedvalue of both humans and AI systems are preserved and enhanced, rather than as a competitionbetween the two.[31]

Estimates suggest that the market for AI in cybersecurity will grow from $3.92 billion in 2017 to $34.81 billion by 2025, at a compound annual growth rate (CAGR) of 31.38% during the forecast period.[32] According to a recent Capgemini survey, the pace of adoption of AI solutionsfor cybersecurity is skyrocketing. The number of companies implementing these systems has risen from one fifth of the overall sample in 2019, to two thirds of companies planning to deploy them in 2020. 73% of the sample tested AI applications in cybersecurity. The most common applications are network security, followed by data security, and endpoint security. Three maincategories can be identified in AI use in cybersecurity: detection (51%), prediction (34%), and response (18%).[33]

The driving forces that are boosting the use of AI in cybersecurity comprise:[34]

1. *Speed of impact:* In some of the major attacks, the average time of impact on organisations is four minutes. Furthermore, today's attacks are not just ransomware, or just targeting certain systems or certain vulnerabilities; they can move and adjust based on what the targets are doing. These kinds of attacks impact incredibly quickly and thereare not many human interactions that can happen in the meantime.

2. *Operational complexity:* Today, the proliferation of cloud computing platforms and the fact that those platforms can be operationalised and deliver services very quickly – in themillisecond range – means that you cannot have a lot of humans in that loop, and you have to think about a more analytics-driven capability.

3. *Skills gaps in cybersecurity remain an ongoing challenge:* According to Frost & Sullivan,[35]there is a global shortage of about a million and a half cybersecurity experts. This level ofscarcity pushes the industry to automate processes at a faster rate.

AI can help security teams in three ways: by improving systems' *robustness*, *response,* and *resilience.* The report defines this as the 3R model.[36] First, AI can improve systems' *robustness*, that is, the ability of a system to maintain its initial assumed stable configuration even when it

---

[31] K. Skapinetz (2018), "Overcome cybersecurity limitations with artificial intelligence", June (www.youtube.com/ watch?time_continue=10&v=-tIPoLin1WY&feature=emb_title).

[32] MarketsandMarkets, "Artificial Intelligence in Cybersecurity Market by Technology Machine Learning, Context Awareness - 2025", MarketsandMarkets (www.marketsandmarkets.com/Market-Reports/ai-in-cybersecurity- market-

224437074.html).

[33] CAP Gemini (2019), "Reinventing Cyber security with Artificial Intelligence. The new frontier in digital security",Research Institute.

[34] This section is taken from McAfee's contribution to the kick-off meeting of the CEPS Task Force.

[35] Frost & Sullivan (2017), "2017 Global Information Security Workforce Study", Center for Cyber Safety and Education.

[36] See M. Taddeo, T. McCutcheon and L. Floridi (2019) on this, "Trusting artificial intelligence in cybersecurity is a double-edged sword", *Nature Machine Intelligence*, November.

processes erroneous inputs, thanks to self-testing and self-healing software. This means that AI systems can be used to improve testing for robustness, delegating to the machines the process of verification and validation. Second, AI can strengthen systems' *resilience*, i.e. the ability of a system to resist and tolerate an attack by facilitating threat and anomaly detection. Third, AI can be used to enhance system *response,* i.e. the capacity of a system to respond autonomously to attacks, to identify vulnerabilities in other machines and to operate strategically by deciding which vulnerability to attack and at which point, and to launch more aggressive counterattacks.

Identifying when to delegate decision-making and response actions to AI and the need of an individual organisation to perform a risk-impact assessment are related. In many cases AI will augment, without replacing, the decision-making of human security analysts and will beintegrated into processes that accelerate response actions.

2.1 System robustness

The need to respond to cyberattacks spurs companies to build systems that are self-learning, i.e., able to establish local context and distinguish rogue from normal behaviour.

Robustness can be understood as the ability of a system to resist perturbations that would fundamentally alter its configuration. In other words, a system is robust when it can continue functioning in the presence of internal or external challenges without changing its original configuration.

Artificial Intelligence for software testing (AIST) is a new area of AI research aiming to design software that can self-test and self-heal. Self-testing refers to "*the ability of a system or component to monitor its dynamically adaptive behaviour and perform runtime testing prior to, or as part of the adaptation process*".[37] Hence, this area of research involves methods of constructing software that it is more amenable to autonomous testing, and knows when to deploy such systems and how to validate their correct behaviour.[38] These systems are able to check and optimise their state continuously and respond quickly to changing conditions. AI- powered behavioural analytics help compare how a system should run with how it is currently running and what the trigger corrections are.[39]

System robustness implies that AI is able to perform anomaly detection and profiling of anything that is generically different. It should be noted, however, that this approach can createa lot of noise from benign detections and false negatives when sophisticated attackers hide byblending in with normal observed behaviours. As such, more robust and accurate approaches focus on detecting attacker's specific and immutable behaviours.

[37] T.M. King et. al. (2019), "AI for testing today and tomorrow: Industry Perspective", IEEE International Conference on Artificial Intelligence Testing, IEEE, pp. 81-88.

[38] See AISTA, Self-Testing (www.aitesting.org/self-testing-ai).

[39] Wired Insider, "Fighting Cybercrime with Self-Healing Machines", *Wired*, October 2018.

System robustness can also be enhanced by incorporating AI in the system's development to increase security controls, for example via vulnerability assessment and scanning. Vulnerabilityassessment can be either manual, assistive, or fully automated. Fully automated vulnerability assessment leverages AI techniques and allows for considerable financial gains and time reductions. ML has been used to build predictive models for vulnerability classification, clustering, and ranking. Support-vector machines (SVMs), Naive Bayes, and Random Forests areamong the most common algorithms. Various evaluation metrics are used to determine the performance, such as precision,[40] recall[41] and f-score.[42] Among other techniques, ML can be used to create risk-analysis models that proactively determine and prioritise security loopholes.[43] Automated planning has also been successfully applied for vulnerability assessment, mainly in the area of generating attack plans that can assess the security of underlying systems. The real-time steps of an attacker are modelled through automated planning, for example by simulating realistic adversary courses of action or focusing on malicious threats represented in the form of attack graphs. Khan and Parkinson suggest that ifattack plans are generated by an AI system, there is greater potential to discover more plans than if they are generated by human experts.[44]

Code review is another area of application for enhancing system robustness. Peer code reviewis a common best practice in software engineering where source code is reviewed manually byone or more peers (reviewers) of the code author. Automating the process by using AI systemscan both reduce time and allow a greater number of bugs to be discovered than ones discovered manually. Several AI systems are being developed for code review support. In June2020, for example, the Amazon Web Services' AI-powered code reviewer from CodeGuru was made publicly available.[45]

The use of AI to improve system robustness not only has a tactical effect (i.e. improving the security of systems and reducing their vulnerability) but also a strategic one. Indeed, it decreases the impact of zero-days attacks. Zero-days attacks leverage vulnerabilities that are exploitable by attackers as long as they remain unknown to the system providers or as long asthere is no patch to resolve them. By decreasing the impact of zero-days attacks, AI reduces their value on the black market.[46]

---

[40] Precision is a metric that quantifies the number of correct positive predictions made.

[41] Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictionsthat could have been made.

[42] F-Measure provides a way to combine both precision and recall into a single measure that captures both properties.

[43] For more on ML techniques for performing fully automated vulnerability assessment, see S. Khan and S. Parkinson (2018), "Review into State of the Art of Vulnerability Assessment using Artificial Intelligence", *Guide to Vulnerability Analysis for Computer Networks and Systems*, Springer, Cham, pp.3-32.

[44] Ibid.

[45] See Amazon, CodeGuru (https://aws.amazon.com/it/codeguru/).

[46] M. Taddeo T. McCutcheon and L. Floridi (2019), "Trusting artificial intelligence in cybersecurity is a double-edged sword", *Nature Machine Intelligence*, November, pp. 1-4.

2.1 System resilience

Resilience can be understood as the ability of a system to resist and tolerate an attack by facilitating threat and anomaly detection. In other words, a system is resilient when it can adaptto internal and external challenges by changing its methods of operations while continuing to function. System resilience implies, unlike system robustness, some fundamental shift in the core activities of the system that has to adapt to the new environment. Threat and anomalies detection (TAD) is today the most common application of AI systems. Indeed:

- There are now approximately 592,145 new unique malware files every day, and possiblyeven more.
- Classification of new threats by humans alone is impossible, and besides, threats are becoming more complicated and better dissimulated.
- In the past, it was common to use signatures to classify malicious attacks, leveraging databases of known threats. Such measures, however, are becoming considerably less effective against the latest strains of advanced malware, which evolve by the second.[47]

AI solutions for cybersecurity enable a fundamental shift from a signature-based detection to a more flexible and continuous monitoring of the network as it shifts from its normal behaviours. *"AI algorithms can detect any changes that appear abnormal – without needing anadvance definition of abnormal."*[48] AI can also provide insights into potential attacks by performing deep packet traces through internal or external sensors or pieces of monitoring software.[49]

Companies use AI to automate cyber defences against spam and phishing and to detect malware, fraudulent payments, and compromised computers and network systems.[50] Furthermore, AI is used for critical forensics and investigative techniques. In particular, AI is used to create real-time, customer-specific analysis, improving the total percentage of malware identified and reducing false positives. Hence, AI data processing helps cybersecurity threat intelligence become more effective. Finally, organisations are using AI-based predictive analytics to determine the probability of attacks, enhancing an organisation's network defence through near real-time data provisions. Predictive analytics can help in processing real-time data from various sources and identifying attack vectors by helping manage big data; in filteringand parsing the data before they are analysed; in automatically filtering out duplicates; in categorising information; and by suggesting which incident to prioritise. In this way predictive analytics reduces human errors and the workload for security analysts.[51]

---

[47] This section is taken from Palo Alto Network's contribution to the fourth meeting of the CEPS Task Force.

[48] R. Goosen et al. (2018), "Artificial intelligence is a threat to cybersecurity. It's also

a solution", The Boston Consulting Group.

[49] Ibid.

[50] Companies like McAfee have access to 1bn sensors via their end points, web gateway, cloud, and CASB protection services and use ML to transform raw data into analytics and insight.

[51] WhoisXML API (2019), "The importance of Predictive Analytics and Machine Learning in Cybersecurity", CircleID, September.

While the use of AI in cybersecurity is increasingly indispensable, AI systems will continue to require a rather collaborative environment between AI and humans, at least for the foreseeablefuture. While completely autonomous systems do exist, their use is as yet relatively limited, and systems still often require human intervention to function as intended.

In this respect, the people involved have to keep monitoring the system (for accuracy, to change request, etc.). Some models still have to be retrained every single day just to stay aheadof the attackers, as attacks change in response to the defences being built. Finally, there are communities of security practitioners that continue to work together to establish a common understanding of what is malicious and what is not.[52]

2.1 System response

System resilience and response are deeply intertwined and logically interdependent, as, to respond to a cyberattack, you need to detect what it is occurring and develop and deploy an appropriate response by deciding which vulnerability to attack and at which point, or by launching counterattacks. During the 2014 Defence Advanced Research Projects Agency (DARPA) Cyber Grand Challenge seven AI systems fought against each other, identifying and patching their own vulnerabilities while exploiting their opponents' flaws without human instructions. Since then, prevention of cyberattacks is increasingly going in the direction of systems able to deploy real-time solutions to security flaws. AI can help to reduce cybersecurityexperts' workloads by prioritising the areas that require greater attention and by automating some of the experts' tasks.[53] This aspect is particularly relevant if one considers the shortage inthe supply of cybersecurity professionals, which is currently estimated at four million workers.[54]

AI can facilitate attack responses by deploying, for example, semi-autonomous lures that create a copy of the environment that the attackers are intending to infiltrate. These deceive them and help understand the payloads (the attack components responsible for executing an activity to harm the target). AI solutions can also segregate networks dynamically to isolate assets in controlled areas of the network or redirect an attack away from valuable data.[55] Furthermore, AI systems are able to generate adaptive honeypots (computer systems intended to mimic likely targets of cyberattacks) and honeytokens (chunks of data that look attractive to potential attackers). Adaptive honeypots are more complex than traditional honeypots insofar as they change their behaviour based on the interaction with attackers. Based on the attacker's reaction to the defences, it is possible to understand its skills and tools. The AI solution gets tolearn the attacker's behaviour via this tool so that it will be recognised and tackled during future attacks.

---

[52] This section is taken from Palo Alto Network's contribution to the fourth meeting of the CEPS Task Force.

[53] R. Goosen et al. (2018), "Artificial intelligence is a threat to cybersecurity. It's also a solution", The Boston Consulting Group.

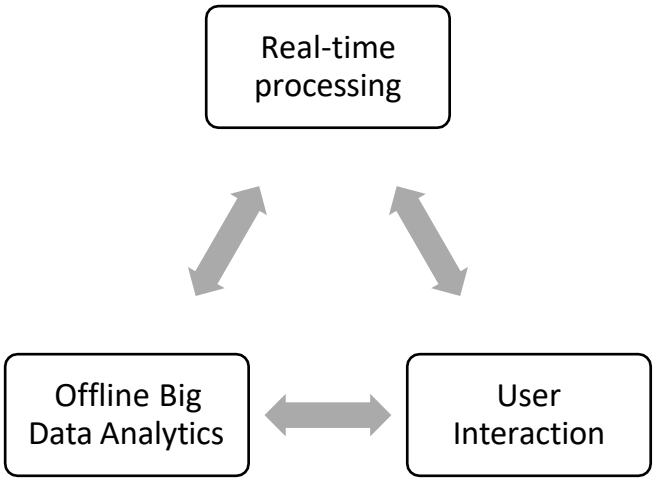
[54] $(ISC)^2$ (2019), "Cybersecurity Workforce Study Strategies for Building and Growing Strong Cybersecurity Teams" (www.isc2.org/-/media/ISC2/Research/2019-Cybersecurity-Workforce-Study/ISC2-Cybersecurity-Workforce- Study-2019.ashx?la=en&hash=1827084508A24DD75C60655E243EAC59ECDD4482).

[55] Ibid.

2.1 Major techniques in the use of AI for system robustness, resilience, and response

Whenever AI is applied to cyber-incident detection and response the problem solving can be roughly divided into three parts, as shown in Figure 2. Data is collected from customer environments and processed by a system that is managed by a security vendor. The detectionsystem flags malicious activity and can be used to activate an action in response.

Figure 2. AI cyber incidents detection and response



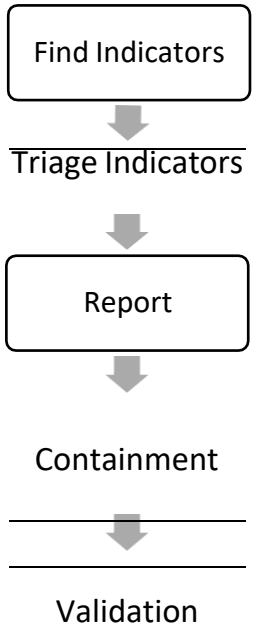
Source: Palo Alto Network contribution to the fourth meeting of the CEPS Task Force.

Companies today recognize that the attack surface is growing massively because of the adoption of the Internet of Things (IoT) and the diffusion of mobile devices, compounded by adiverse and ever-changing threat landscape. Against this backdrop, there are two measures that can be implemented:

1. speed up defenders
2. slow down attackers.

With respect to speeding up defenders, companies adopt AI solutions to automate the detection and response to attacks already active inside the organization's defenses. Security teams traditionally spend a lot of time dealing with alerts, investigating if they are benign or malicious, reporting on them, containing them, and validating the containment actions. AI canhelp with some of the tasks that security operations teams spend most of their time on. Notably, this is also one of the primary and most common uses of AI in general.

In particular, security operations teams can use AI to solve the following five fundamental questions:



1. Classify items: What does this item represent? How mature is it?
2. What is the scale of it?
3. How is it organized?
4. Is it anomalous? Is it a behavior an active attacker would manifest?
5. What actions should be taken?

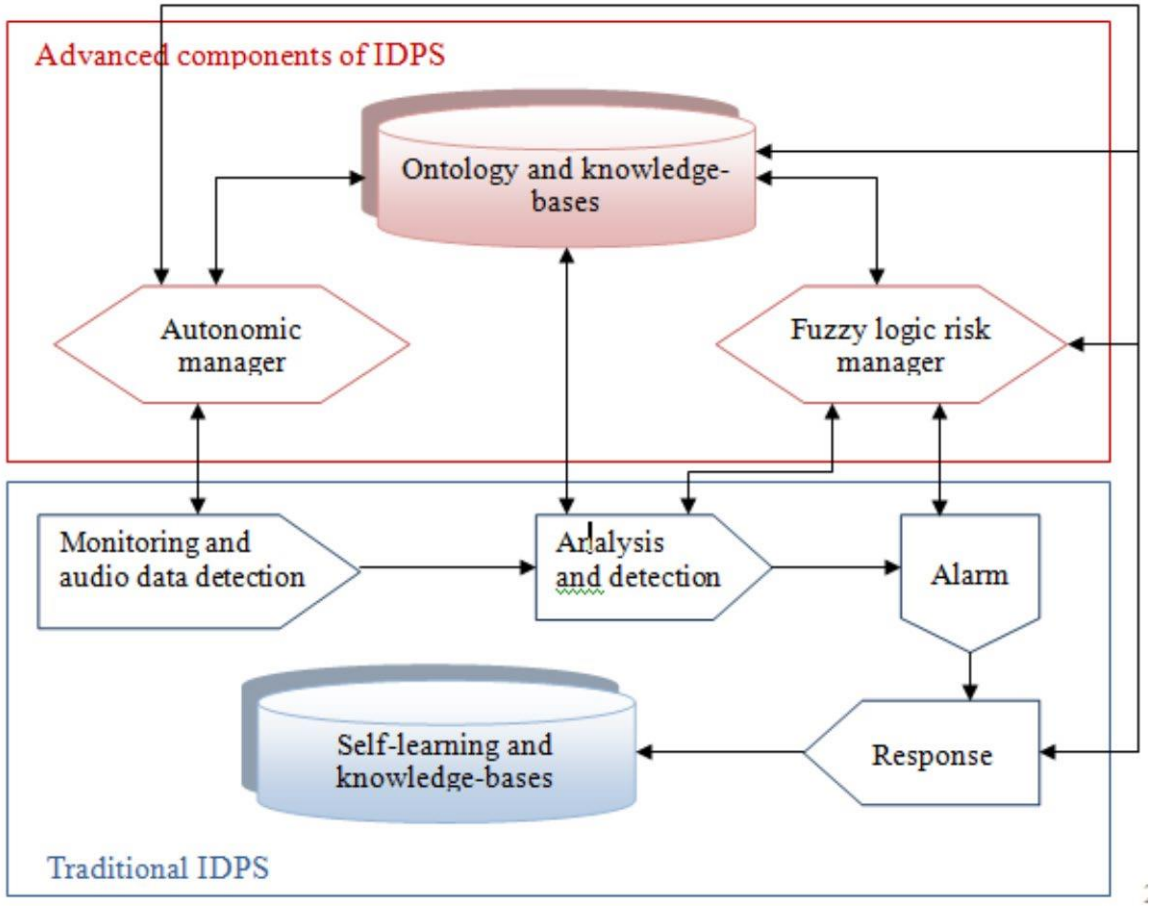
In the past, the industry focused first on labelling and on categorizing malware, but nowadays companies are using models that are not looking for individual pieces of malware; rather they are looking at howattackers behave. It is therefore becoming more common to use ML threat-detection models that are behavioral in their analysis, and in turn they are becoming durable and potentially capable of detecting zero-day attacks. The goal is to identify the very subtle yet immutable attack behavior with high fidelity, and low noise.

The following are practical examples of the benefits of using AI and ML for cybersecuritydetection and response.[56]

- ML trained on user interaction provides a way of understanding local context and knowing what data to focus on; models trained to identify those more likely to be malicious improve the efficiency of a system by triaging the information to process in real time. In this way, using ML is cost saving but also allows for faster reaction in the most critical situations.
- ML can be useful in detecting new anomalies by learning robust models from the data they have been fed with. ML is particularly good at identifying patterns and extracting algorithms in large sets of data where humans are lost.
- ML can be useful for asynchronous user profiling and for measuring deviation from common behaviors as well as going back to much larger data volumes to understand behavior.
- ML trained on immutable attacker 'Tactics, Techniques, and Procedures' (TTP) behaviors (those identified in the Mire Attack framework)[57] can support durable and broad attacker detection.

To better illustrate the use of AI and ML for cybersecurity detection and response, Figure 3 presents an intrusion detection and prevention system that combines software and hardware devices inside the network. The system *"can detect possible intrusions and attempt to prevent them. Intrusion detection and prevention systems provide four vital security functions: monitoring, detecting, analyzing and responding to unauthorized activities."*[58]

Figure 3. Intrusion detection and prevention system



Source: Dilek (2015).

There are a variety of AI techniques that can be used for intrusion prevention, detection, and response. Table 1 illustrates examples of the main advantages of some of these techniques.[59]

Table 1. Examples of AI techniques for intrusion prevention, detection and response

Technology	Advantages
Artificial Neural Networks [60]	Parallelism in information processingLearning by example Nonlinearity – handling complex nonlinear functionsResilience to noise and incomplete data Versatility and flexibility with learning models
Intelligent Agents[61]	Mobility Rationality – in achieving their objectives Adaptability – to the environment and user preferences Collaboration – awareness that a human user can make mistakes and provide uncertain or omit important information; thus, they should not accept instructions without consideration and checking the inconsistencies with the user
Genetic Algorithms[62]	Robustness Adaptability to the environment Optimization – providing optimal solutions even for complex computing problemsParallelism – allowing evaluation of multiple schemas at once Flexible and robust global search
Fuzzy Sets[63]	Robustness of their interpolative reasoning mechanismInteroperability – human friendliness

Source: Dilek (2015).

All these intrusion detection AI-powered technologies help in reducing the dwell time – the length of time a cyberattacker has free reign in an environment from the time they get in untilthey are eradicated.[64] In December 2019, the dwell time in Europe was about 177 days, and attackers were discovered in only 44% of cases because of data breach or other problems. Using AI techniques, the dwell time has been dramatically reduced.[65]

Finally, AI can be also very helpful in enhancing network security. (See Box 1).

[60] First developed in 1957 by Frank Rosenblatt, these techniques rely on the perceptron. By connecting with one another and processing raw data, perceptrons independently learn to identify the entity on which they have been trained. See A. Panimalar et al. (2018), "Artificial intelligence techniques for cybersecurity", *International Research Journal of Engineering and Technology* (IRJET), Vol. 5, No. 3.

[61] Intelligent Agents are defined as entries able to recognize movement through their sensors, to follow up on an environment based on the perceived condition using

actuators and to direct their behaviour toward the accomplishment of an objective. They can vary greatly in complexities (thermostats, for example, are intelligent agents). In cybersecurity, they can be used in showdown DDoS attacks, and could potentially be deployed as Cyber Police mobile agents. See A. Panimalar et al. (2018), op. cit.

[62] The genetic algorithm is a method for solving both constrained and unconstrained optimisation problems that is based on natural selection, the process that drives biological evolution.

[63] Fuzzy sets can be considered an extension and simplification of classical sets. They can be understood in the context of set membership. They allow partial membership of elements that have varying degrees of membership in the set.

[64] See Optiv, "Cybersecurity Dictionary, Dwell Time" (www.optiv.com/cybersecurity-dictionary/dwell-time).

[65] M. Walmsley (2019), intervention at the CEPS Cyber Summit 2019, December (www.youtube.com/ watch?v=sY16ToU9UiQ [3:05:40]).

Box 1. AI and network security

Example 1. Detecting route hijacking attacks[66]

AI is helpful in enhancing network security. An increasingly popular cyberattack today is hijackingInternet Protocol (IP) addresses. 'Route hijacking' means stealing traffic intended for other destinations. The regions of the Internet in the world are connected through a global routing protocol called the Border Gateway Protocol (BGP), which allows different parts of the Internet to talk to each other. Using the BGP, networks exchange routing information in such way that packets are able to reach the correct destination. Each region announces to its neighbourhood that it holds certain IP addresses. There are about 70,000 regions on the Internet called autonomous systems and about 700,000 distinct announcements. The BGP does not have any security procedures for validating that a message is actually coming from the place it says it's coming from, so hijackers exploit this shortcoming by convincing nearby networks that the best way to reach a specific IP address is through their network. In other words, a rogue region can announce that it contains an IP address that belongs, for instance, to MIT. A malicious router would be advertising a network that does not really belong to its autonomous system (the rangeof IP addresses that it has authority over). In so doing, the malicious router and related infrastructure can eavesdrop, and redirects the traffic that was supposed to go to MIT to the rogue region. This is happening regularly, for example to send spam and malware or when an actor manages to hijack bitcoin traffic to steal the bitcoins.

In a recent joint project between MIT and the University of California at San Diego, researchers have trained a machine-learning model to automatically identify malicious actors through the patterns of their past traffic. Using data from network operator mailing lists and historical BGP data, taken every five minutes from the global routing tables during a five-year period, the machine-learning model was able to identify malicious actors. Their networks had key characteristics related to the specific blocks of IP addresses they use, namely:

- Volatile changes in activity: if a region announces address blocks and then the announcements disappear in a short time, the likelihood of there being a hijacker becomes very high. The average duration of an announcement for legitimate networks was two years,compared with 50 days for hijackers.
- Multiple address blocks: serial hijackers advertise many more blocks of IP addresses.
- IP addresses in multiple countries: most networks do not have foreign IP addresses, but hijackers are much more likely to announce addresses registered in different countries and continents.

One challenge in developing this system was handling the false positives related

to a legitimate short-term address announcement or human error. Indeed, changing the route is sometimes a legitimate way to block an attack.

This model allows network operators to handle these accidents in advance by tracing hijackers' behaviour instead of reacting on a case-by-case basis.

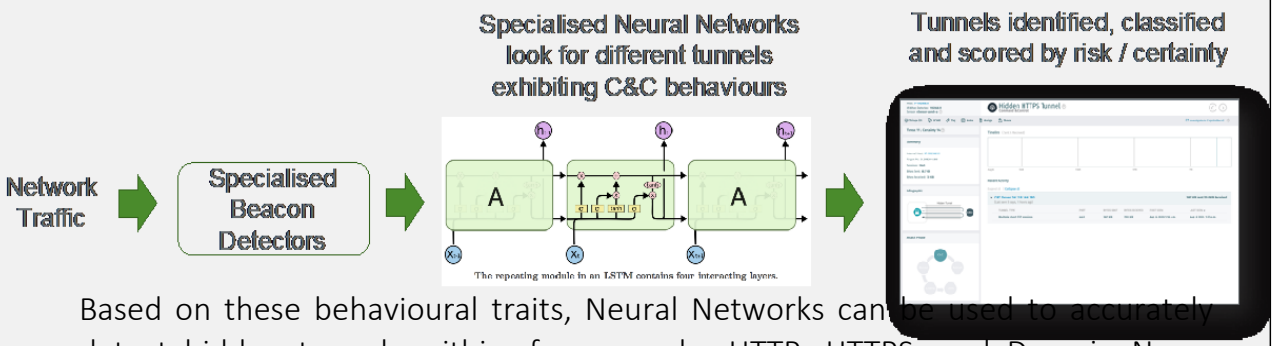
[66] This section draws from the intervention of Professor David Clark from MIT at the third meeting of the CEPS Task Force and from A. Conner-Simons (2019), "Using machine learning to hunt down cybercriminals", MIT CSAIL, October.

The MIT model is particularly relevant when considering more generally that the Internet was not designed as a high-security network. Incremental security improvements primarily address specific attacks, but these might fail to solve the fundamental problems and could also introduce new undesirable consequences (e.g., Border Gateway Protocol Security prevents route hijacking but causes delayed route convergence, and does not support prefix aggregation, which results in reduced scalability).[i]

Example 2. Detecting hidden tunnel attacks[ii]

Identifying attackers who are already operating inside compromised networks is a more complex challenge. Sophisticated attackers use hidden tunnels to carry out command-and-control and exfiltration behaviours. This means that they steal critical data and personally identifiable information (PII) by blending in with normal traffic, remotely controlling the theft of information, and slipping it out through those same tunnels. Because they blend in with multiple connections that use normal, commonly allowed protocols, hidden tunnels are very difficult to detect.

AI can constantly perform a highly sophisticated analysis of metadata from network traffic, revealing subtle abnormalities within a protocol that gives away the presence of a hidden tunnel. Even though messages are disguised within an allowed protocol, the resulting communications that make up the hidden tunnel can't help but introduce subtle attack behaviours into the overall conversation flow. These include slight delays or abnormal patterns in requests and responses.



Based on these behavioural traits, Neural Networks can be used to accurately detect hidden tunnels within, for example, HTTP, HTTPS, and Domain Name System (DNS) traffic without performing any decryption or inspection of private payload data. It doesn't matter what field attackers use to embed communications or whether they use a never-before-seen obfuscation technique. The attacker's variance from normal protocol behaviour will still expose the hidden tunnel's presence to the Neural Networks.

[i] While the contribution of AI/ML to cybersecurity is of relevance, it is critical that cybersecurity be addressed at the root wherever possible. Scalability, Control and Isolation on Next Generation Networks (SCION) is an Internet- compatible (IPv4 and IPv6) architecture that addresses today's network security issues on the

Internet (www.scion- architecture.net).

[ii] See "Breaking ground: Understanding and identifying hidden tunnels" (www.vectra.ai/blogpost/breaking-ground- understanding-and-identifying-hidden-tunnel).

3. AI malicious uses

AI developments bring not only extensive possibilities, but also many corresponding challenges. People can use AI to achieve both honourable and malevolent goals.

The impact of AI on cybersecurity is usually described in terms of expanding the threat landscape. The categories of actors and individuals enabled through AI to carry out malicious attacks are proliferating. At the same time, new forms of attacks against AI systems – different in nature from traditional cyberattacks – increase the attack surface of connected systems in an exponential and sometimes unmeasurable way.

As far as these shifts are concerned, researchers agree that AI affects the cybersecurity landscape by:

- expanding existing threats
- introducing new threats
- altering the typical characteristics of threats.[67]

3.1 Expansion of existing threats

The availability of cheap and increasingly effective AI systems for attacks means categories of individuals and groups have the potential to become malicious actors. This means the asymmetry that once existed in the power balance between conventional and unconventional actors is increasingly shrinking. With the widening spectrum of actors capable of meaningfully undertaking a potentially significant attack, such as those against critical infrastructures, the malicious use of AI applications has become one of the most discussed aspects of this technology.

Experts refer to this phenomenon as the 'democratisation of artificial intelligence', meaning both the increasing number of potential actors exploiting AI to perform an attack, and the democratisation of the software and AI systems themselves. Indeed, the ease of access to scientific and engineering works around machine learning partly explains the increasing availability of AI to a greater number of individuals.[68] In modern times, access to software codes has become an increasingly easy task. Open repositories of stored software programming allow anyone with a laptop and the discrete knowledge to be able to explore the source code of a lot of software, including AI. This is even more relevant in a context in which there is already wide disclosure of hacking tools. Furthermore, academic and scientific research on AI is often openly disseminated, and made available to the general public with little review of misuse-prevention measures, and even fewer boundaries[69] on the vulgarisation of such outcomes. The issue of research openness will be further explored in this report.

[67] See M. Brundage et al. (2018), "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation", Malicious AI Report, February, p. 18.

[68] As J.-M. Rickli puts it, "*artificial intelligence relies on algorithms that are easily replicable and therefore facilitate proliferation. While developing the algorithm takes some time, once it is operational, it can be very quickly and easily copied and replicated as algorithms are lines of code*", J.-M. Rickli (2018), "The impact of autonomy and artificial intelligence on strategic stability", UN Special, July-August, pp. 32-33.

[69] For instance, *"(…) it is generally much easier to gain access to software and relevant scientific findings. Indeed, many new AI algorithms are reproduced in a matter of days or weeks. In addition, the culture of AI research is characterized by a high degree of openness, with many papers being accompanied by source code.",* M. Brundage (2018), op.cit., p. 17.

The automation of tasks previously undertaken by humans is another effect of the democratisation of AI. As Ferguson puts it, "*Imagine your attacker has the ability to conduct real-time impersonation of both systems and people, no longer harvesting passwords with noisy pen-testing tools, but through real-time adaptive shimming of the very systems it seeks later toexploit.*"[70] As more and more people use ML, the pattern of time-consuming tasks could be speeded up, rendering them more effective, and making cyber capabilities that were once the preserve of large industry players or wealthy governments accessible to small groups and individuals.[71]

The cost-availability nexus is another factor in the democratisation of AI that leads to the widening spectrum of malicious actors. As Comiter points out: *"the proliferation of powerful yet cheap computing hardware means almost everyone has the power to run these algorithms on their laptops or gaming computers. […] it does have significant bearing on the ability for non- state actors and rogue individuals to execute AI attacks. In conjunction with apps that could be made to allow for the automation of AI attack crafting, the availability of cheap computing hardware removes the last barrier from successful and easy execution of these AI attacks.*"[72]

To sum up, the spectrum of malicious actors is being widened by the proliferation of cheap computing hardware, the growing availability and decreasing cost of computing capability through the cloud, and the open-source availability of most of the tools that could facilitate model training and potentially malicious activities.

The greater accessibility of AI tools also affects the combination of efficiency and scalability.[73] Some of the AI systems that are replacing tasks once assigned to humans are destined to departfrom ordinary human performance. They will run in a faster way, and will execute those tasks a greater number of times.[74] In the cybersecurity context, scalability will allow an attack to reproduce at a level that has not been seen before. By using the example of spear-phishing attacks, Brundage et al point to two basic effects of scalability and efficiency for the actors driving an attack with an AI system.[75] On the one hand, cheap and efficient AI systems will, as mentioned, expand the category of adversaries being able to handily access such applications.On the other hand, actors that were already present in the threat landscape and labelled as

---

[70] R. Ferguson (2019), "Autonomous Cyber Weapons - The Future of Crime?", *Forbes*, 10 September (www.forbes.com/sites/rikferguson1/2019/09/10/autonomous-cyber-weapons-the-future-of-crime/#549591f85b1a).

[71] M.C Horowitz et al. give the example of the 'script kiddies', i.e. "…*relatively unsophisticated programmers, (…) who are not skilled enough to develop their own cyber-attack programs but can effectively mix, match, and executecode developed by others? Narrow AI will increase the capabilities available to such actors, lowering the bar for attacks by individuals and non-state groups and increasing the scale of potential attacks for all actors*.", M.C Horowitz et al. (2018), "Artificial Intelligence and International Security", Center for a New American Security, p. 13.

[72] M. Comiter (2019), "Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It", Belfer Center for Science and International Affairs, Harvard Kennedy School, August, p. 53.

[73] OECD (2019a), op. cit., p. 96.

[74] See M. Brundage et al. (2018), op. cit., p. 5 and p. 16. Nonetheless, the devolution of tasks from humans to machines do encounter a certain limits. For instance, see B. Buchanan and T. Miller (2017), "Machine Learning for Policymakers", Belfer Center for Science and International Affairs, Harvard Kennedy School, p. 20; See also K. Grace et al. (2017)*, When Will AI Exceed Human Performance? Evidence from AI Experts*, ArXiv.

[75] See also OECD (2019a), op. cit., p. 96.

potential malicious attackers will be able to benefit from AI systems to which they already hadaccess, with a much higher efficiency rate.[76]

The wider distribution of AI systems not only multiplies the opportunities for cyberattacks – byincreasing their speed and volume – but also allows them to become more sophisticated, for example by making their attribution and detection harder. AI also allows for the discovery of flaws that were never discovered before. Attackers, for instance, are able to more easily discover vulnerabilities generating new payloads fuzzing to discover new issues. Unusualbehaviour triggers abnormal responses in the system, and AI systems, trained by already- discovered payloads for existing vulnerabilities, can suggest new payloads that would increasethe chances of discovering new systems' exposures. AI can also help to exploit, not just discover, these newly discovered vulnerabilities by generating exploit variants and running them faster.[77]

Finally, it appears that such an increase of actors also impacts national and international security, particularly because of the inherent dual use of AI technology. According to the available literature, the repurposing of easily accessible AI systems is already having a significant effect on the development of lethal autonomous weapons systems (LAWS).[78] The availability of accessible AI solutions will also expand the possibility of warfare activities and tasks that will have a strategic impact being relayed to surrogates to conduct. Both state and non-state actors are increasingly relying on technological surrogates such as AI to be used as aforce multiplier. An example of this is the alleged meddling in the 2016 US election, when a disinformation campaign aimed to persuade targeted voters to support the winning candidate.Another example is the US offensive operation carried out in 2019 as part of the ongoing cyberwar against Iran. This disabled a critical database that Iran was using to plot attacks againstUS oil tankers.[79]

3.1.1 Characteristics of AI-powered attacks80

Three characteristics of AI are likely to affect the way in which AI-powered attacks are carried out:

1. *Evasiveness*: AI is helping to modify the way in which attacks are detected. An AI-powered malware is much more difficult to detect by an anti-malware. The case in point

---

[76] M. Brundage et al. (2018), op. cit., p. 18.

[77] I. Novikov (2018), "How AI Can Be Applied To Cyberattacks", *Forbes*, 22 March (www.forbes.com/sites/ forbestechcouncil/2018/03/22/how-ai-can-be-applied-to-cyberattacks/#27ef6e9849e3).

[78] C. Czosseck, E. Tyugu and T. Wingfield (eds), (2011), "Artificial Intelligence in Cyber Defense", Cooperative Cyber Defense Center of Excellence (CCD COE) and Estonian Academy of Sciences, 3rd International Conference on Cyber Conflict, Tallinn, Estonia. According to Ferguson, "*The repurposing of this technology will undoubtedly start at the*

nation-state level, and just like everything else it will trickle down into general availability. It is already past time for defenders to take the concept of autonomous cyber weapons seriously." Ferguson, R (2019), "Autonomous Cyber Weapons - The Future of Crime?" *Forbes*, 10 September (www.forbes.com/sites/rikferguson1/2019/09/10/autonomous-cyber-weapons-the-future-of-crime/#549591f85b1a).

[79] See J. E. Barnes (2019), "Hurt Iran's Ability to Target Oil Tankers, Official Says", *New York Times*, 28 August (www.nytimes.com/2019/08/28/us/politics/us-iran-cyber-attack.html).

[80] This section draws from the contribution of Marc Ph. Stoecklin, IBM Research Centre Zurich and member of the Task Force Advisory Board, in the kick-off meeting of the CEPS Task Force.

is represented by IBM's Deeplocker malware. This is a new class of highly targeted malware that uses AI to hide its nature in benign applications, such as video conferencing applications, and identifies its target through face recognition, voice recognition or geo-localisation. The malware can conceal its intent until it reaches the defined target, which makes it fundamentally different from the classic 'spray-and-pray' attacks.

1. *Pervasiveness*: On 13 March 2004 driverless cars competed in the DARPA Grand Challenge in the Mojave Desert. Although this was deemed a failure because no vehicle achieved anything close to the goal, the improvements in driverless car technology have since been enormous. In 2016 DARPA launched the Cyber Grand Challenge in which competitors were asked to bring bots able to compete against each other without human instructions. As with the self-driving vehicles, the future pervasive potential of these new technologies is clear. This era of pervasive intelligence will be marked by a proliferation of AI-powered smart devices able to recognise and react to sights, sounds, and other patterns. Machines will increasingly learn from experience, adapt to changing situations, and predict outcomes. The global artificial intelligence market size was valued at $39.9 billion in 2019 and is expected to grow at a CAGR of 42.2% from 2020 to 2027.[81]

2. *Adaptiveness*: AI is adaptive, meaning that it can learn and to some extent become creative, and come up with ideas that attackers would not necessarily have thought of. During the DEF CON Hacking conference in 2017, a group of researchers showed how they successfully attacked a web application through an AI that found its way in using the Structured Query Language (SQL) database injection attack. The distinctiveness of this attack was that the AI figured out by itself how the SQL injection worked.

3.2 Introduction of new threats

As well as existing threats expanding in scale and scope, progress in AI means completely new threats could be introduced. The AI characteristics of being unbounded by human capabilities could allow actors to execute attacks that would not otherwise be feasible.

3.2.1 Deepfakes[82]

The use of 'deepfakes' has been steadily rising since a Reddit user first coined the term in 2017. Deepfakes are a developing technology that use deep learning to make images, videos, or texts of fake events. There are two main methods to make deepfakes. The first is usually adopted for 'face-swapping' (i.e., placing one person's face onto someone else's), and requires thousands

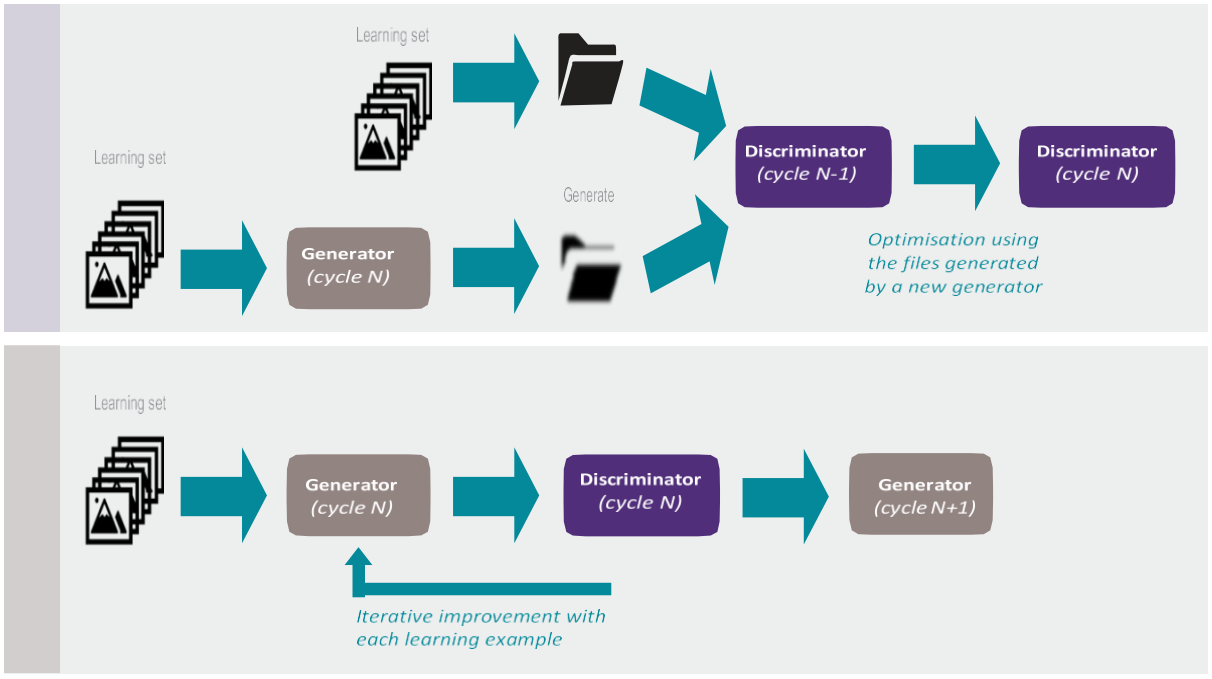
[81] Grand View Research (2020), "Artificial Intelligence Market Size, Share & Trends Analysis Report By Solution (Hardware, Software, Services), By Technology (Deep Learning, Machine Learning), By End Use, By Region, And Segment Forecasts, 2020 – 2027", July.

[82] This section of the report was contributed by Jean-Marc Rickli from the Geneva Centre for Security Policy (GCSP) and member of the Advisory Board of the Task Force, with the help of Alexander Jahns.

of face shots of the two people to be run through an AI algorithm called an encoder. The encoder then finds and learns similarities between the two faces, and reduces them to their shared common features, compressing the images in the process. A second AI algorithm calleda decoder is then taught to recover the faces from the compressed images: one decoder recovers the first person's face, and another recovers the second person's face. Then, by givingencoded images to the 'wrong' decoder, the face-swap is performed on as many frames of a video as possible to make a convincing deepfake.[83]

The second and very important method to make deepfakes is called a generative adversarial network (GAN). A GAN pits two AI algorithms against each other to create brand new images (see Figure 4). One algorithm, the generator, is fed with random data and generates a new image. The second algorithm, the discriminator, checks the image and data to see if it corresponds with known data (i.e. known images or faces). This battle between the two algorithms essentially winds up forcing the generator into creating extremely realistic images (e.g. of celebrities) that attempt to fool the discriminator.[84]

Figure 4. The functioning of a generative adversarial network



Source: C. Meziat and L. Guille (2019), "Artificial Intelligence and Cybersecurity", Wavestone, 5 December.

[83] I. Sample (2020), "What are deepfakes and how can you spot them" *The Guardian*, 13 January (www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them).

[84] K. Vyas (2019), "Generative Adversarial Networks: The Tech Behind DeepFake and FaceApp", *Interesting Engineering*, 12 August (https://interestingengineering.com/generative-adversarial-networks-the-tech-behind-deepfake-and-faceapp).

These images have been used to create fake yet realistic images of people, with often harmful consequences. For example, a McAfee team used a GAN to fool a facial recognition system like those currently in use for passport verification at airports. McAfee relied on state-of-the-art, open-source facial-recognition algorithms, usually quite similar to one another, thereby raisingimportant concerns about the security of facial-recognition systems.[85]

Deepfake applications also include text and voice manipulation as well as videos. As far as voicemanipulation is concerned, Lyrebird claims that, using AI, it was able to recreate any voice usingjust one minute of sample audio, while Baidu's Deep Voice clones speech with less than four seconds of training. In March 2019, AI-based software was used to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000). In this case, the CEO thought he was talking to the chief executive of the firm's German parent company, who demanded the payment be made to a Hungarian subsidiary.

Deepfakes used for text manipulation is also increasingly concerning. With GPT-3 generative writing it is possible to synthetically reproduce human-sounding sentences that are potentiallyeven more difficult to distinguish from human-generated ones than video content. Even with state-of-the art technology, it is still possible to tell video content has been synthetically produced, for example from a person's facial movements being slightly off. But with GPT-3 output there is no unaltered original that could be used for comparison or as evidence for a fact check.[86] Text manipulation has been used extensively for AI-generated comments and tweets. Diresta highlights how *"seeing a lot of people express the same point of view, often at the same time or in the same place, can convince observers that everyone feels a certain way, regardless of whether the people speaking are truly representative – or even real. In psychology, this is called the majority illusion."* As such, by potentially manufacturing a majority opinion, text manipulation is and will increasingly be applied to campaigns aiming to influence public opinion. The strategic and political consequences are clear.[87]

The malicious use of deepfakes is trending in many areas, as discussed below.

Pornographic

The number of deepfake videos online amounted to 14,678 in September 2019, according to Deeptrace Labs, an 100% increase since December 2018. The majority of these (96%) are pornographic in content, although other forms have also gained popularity.[88] Deepfake technology can put women or men in a sex act without their consent, while also removing theoriginal actor, creating a powerful weapon for harm or abuse. According to a Data and Society

[85] K. Hao and P. Howell O'Neill (2020), "The hack that could make face recognition think someone else is you", *MIT Technology Review*, 5 August

(www.technologyreview.com/2020/08/05/1006008/ai-face-recognition-hack-misidentifies-person/).

[86] R. Diresta (2020), "AI-Generated Text Is the Scariest Deepfake of All", *Wired*, 31 July (www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/).

[87] Ibid.

[88] H. Ajde, G. Patrini, F. Cavalli and L. Cullen (2019), "The State of Deepfakes: Landscape, Threats, and Impact", DeepTrace, September, p. 1 (https://regmedia.co.uk/2019/10/08/deepfake_report.pdf).

report,[89] deepfakes and other audio and visual manipulation can be used with pornography toenact vendettas, blackmail people or trick them into participating in personalised financial scams. The increasing accessibility of this technology makes this even more problematic .[90] Onerecent example is the conjunction of 3D-generated porn and deepfakes, which allow a user toput a real person's face on another person's body, and do whatever violent or sexual act they want with them.[91] Notably, audiovisual manipulation and other less sophisticated methods such as basic video and photo-editing software (part of what Paris and Donovan call 'Cheap Fakes' or 'Shallowfakes'),[92] can also change audiovisual manipulation in malicious ways, much more easily and cheaply.[93]

Political

Deepfakes and cheap fakes also have malicious uses in political settings. Videos of heads of states saying things contrary to common belief have emerged in the past couple of years because of these technologies, and they are less and less easily differentiated from authentic videos. A recent example was when the incumbent UK Prime Minister, Conservative Boris Johnson, appeared to endorse his Labour Party rival, Jeremy Corbyn, and vice versa.[94] While those aware of the political context will see through this hoax, people less aware might believethem completely, and confusion and disorder is created in an important democratic process. This effect can be further maliciously exploited in countries where people have less digital literacy, and even more so as these technologies become more widely usable. In such a context comes Facebook's announcement that the company will remove misleading manipulated media whenever those *"have been edited or synthesized in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say",* and they are *"the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic."*[95]

While deepfake videos could be spotted as such by countering software, they can still proliferate across social media networks in very little time, changing the course of a democraticelection or even just one person's career.[96] Importantly, deepfakes can also be used as a

---

[89] B. Paris and J. Donovan (2019), "DeepFakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence",Data and Society, September, p. 41 (https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal.pdf).

[90] B. Paris and J. Donovan (2019), op. cit., p. 41.

[91] S. Cole and E. Maiberg, (2019), "Deepfake Porn Is Evolving to Give People Total Control Over Women's Bodies", *VICE*, 6 December (www.vice.com/en_us/article/9keen8/deepfake-porn-is-evolving-to-give-people-total-control-over-womens-bodies).

[92] H. Ajde, G. Patrini, F. Cavalli and L. Cullen (2019), op. cit., p. 1.

[93] Ibid., pp. 5-6

[94] S. Cole (2019), "Deepfake of Boris Johnson Wants to Warn You About Deepfakes", *VICE*, 13 November (www.vice.com/en_uk/article/8xwjkp/deepfake-of-boris-johnson-wants-to-warn-you-about-deepfakes).

[95] M. Bickert (2020), "Enforcing Against Manipulated Media", Facebook, 6 January (https://about.fb.com/news/ 2020/01/enforcing-against-manipulated-media/).

[96] A. Ridgway (2021), "Deepfakes: the fight against this dangerous use of AI", Science Focus, 12 November (www.sciencefocus.com/future-technology/the-fight-against-deepfake/).

scapegoat, often in political contexts, with people claiming that harmful video evidence has been altered when it has not. An example of this arose in 2018 when a married Brazilian politician claimed that a video allegedly showing him at an orgy was a deepfake, yet no one hasbeen able to prove it so.[97] Similarly, when the Gabonese president Ali Bongo appeared on camera in a New Year's address at the end of 2018 to end speculation about his health, his political rivals claimed it was a deepfake. Yet experts have been unable to prove this.[98] Consequently, everyday voters and consumers of media need to be aware of the political impact of deepfakes as much as experts and politicians because very convincing fake videos can undermine trust in real ones in the eyes of the public.

Crime and cybersecurity

In recent years criminals have also made malicious use of deepfake technology for financial gain. According to Deeptrace, "Deep fakes do pose a risk to politics in terms of fake media appearing to be real, but right now the more tangible threat is how the idea of deep fakes can be invoked to make the real appear fake. The hype and rather sensational coverage speculatingon deep fakes' political impact has overshadowed the real cases where deep fakes have had animpact", such as cybercrime.[99] While internet and email scams have been around for decades,the advance of deepfake technology in sound and video has allowed for even more intricate and hard-to-spot fraudulent criminal activity.

These sorts of crimes could range from a basic level of hacktivists making false claims and statements to undermine and destabilise a company, to more serious efforts such senior executives confessing to financial crimes or other offences. Deepfakes can also use social engineering to make frauds more credible by using video or audio of, for instance, a member of the targeted organisation, increasing the chances of the attacks succeeding.[100] Market Research company Forrester has claimed that deepfakes could end up costing businesses as much as $250 million in 2020.[101] Software tools that can spot criminal deepfakes are being developed, but it only takes one individual in a company to believe in a modified audio or visualsource for a large amount of damage to be done.

Military

Concern about deepfakes has also reached hard security, with many of the world's militaries nowbeing very worried about them. In 2018, funding began for a US DARPA project that will try to determine whether AI-generated images and audio are distinguishable, using both technological

[97] D. Thomas (2020), "Deepfakes, a Threat to Democracy or Just a Bit of Fun?", *BBC News*, 23 January (www.bbc.com/news/business-51204954).

[98]H. Ajde, G. Patrini, F. Cavalli and L. Cullen (2019), op. cit., p. 10.

[99] Orange Business Services (2020), "Fake news: What could deepfakes and AI scams

mean for cybersecurity?", Orange, 2 January (www.orange-business.com/en/magazine/fake-news-what-could-deepfakes-and-ai-scams- mean-cybersecurity).

[100] C. Meziat et al. (2020), "Deep Dive into Deepfake – how to face increasingly believable fake news? (1/2)", Wavestone, 5 May (www.riskinsight-wavestone.com/en/2020/05/deep-dive-into-deepfake-how-to-face- increasingly-believable-fake-news-1-2/).

[101] Orange Business Services (2020), op. cit.

and non-technological means.[102] Legal witnesses and AI experts told US lawmakers in June 2019that they needed to act immediately to stay ahead of the threat of deepfakes and other AI-ledpropaganda, which could be deployed by adversaries such as Russia ahead of the next presidential election. These efforts are ongoing, with the US Congress greenlighting a $5 millionprogramme to boost new technologies in detecting deepfakes. This reveals that the Pentagon views audiovisual manipulation as a key national security issue. Examples of potential problemsinclude a national security leader giving false orders or acting unprofessionally, which could causechaos.[103] Todd Myers, automation lead for the CIO-Technology Directorate at the National Geospatial-Intelligence Agency, believes that China is the main proactive user of deepfake technology for military reasons, specifically by creating fake bridges in satellite images. "From atactical perspective or mission planning, you train your forces to go a certain route, toward a bridge, but it's not there. Then there's a big surprise waiting for you," he warns.[104]

Finally, as mentioned throughout this analysis, one of the greatest threats of deepfakes to bothpublic and private life is not the technology itself but to its potential to converge with other technologies and bring about new and unexpected challenges. By compounding different technologies, state and non-state actors will be able to further propagate misleading or false narratives, targeting harmful and disruptive content at specific populations with deepfake, IoT, and AI capabilities.[105] It is difficult to predict exactly how this issue of convergence will pan out, but it leaves a lot of room for devastating malicious uses should governments, private companies, and individuals fail to educate and prepare themselves against such threats.

3.1.1 Breaking CAPTCHAs

Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs) were created to preclude automatised programs from being malicious on the world wide web (filling out online forms, accessing restricted files, accessing a website an incredible number oftimes, etc.) by confirming that the end-user is in fact 'human' and not a bot. Today, machine learning is able to break CAPTCHAs in 0.05 seconds, using GAN. Indeed, synthesised CAPTCHAs can be created, along a small dataset of real CAPTCHAs, to create an extremely fast and accurate CAPTCHA solver.[106]

---

[102] W. Knight (2018), "The US military is funding an effort to catch deepfakes and other AI trickery", *MIT Technology Review*, 23 May (www.technologyreview.com/2018/05/23/142770/the-us-military-is-funding-an-effort-to-catch- deepfakes-and-other-ai-trickery/).

[103] J. Keller (2020), "U.S. intelligence researchers eye $5 million program to encourage new technologies in detecting deepfakes", *Military and Aerospace Electronics*, 8

January.

[104] P. Tucker (2019), "The Newest AI-Enabled Weapon: 'Deep Faking' Photos of the Earth", Defense One, 31 March (www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/).

[105] M. Erfourth and A. Bazin (2020), "Extremism on the Horizon: The Challenges of VEO Innovation", Mad Scientist Laboratory, 19 March (https://madsciblog.tradoc.army.mil/220-extremism-on-the-horizon-the-challenges-of-veo-innovation/).

[106] R. Irioni (2018), "Breaking CAPTCHA using Machine Learning in 0,05 Seconds", Medium, 19 December (https://medium.com/towards-artificial-intelligence/breaking-captcha-using-machine-learning-in-0-05-seconds- 9feefb997694) and E. Zouave et al. (2000), "Artificial Intelligence Cyberattacks", FOI, p. 24.

3.1.1 Swarming attacks

AI systems could be used to control robots and malware behaviour that would be impossible for humans to do manually. This could allow 'swarming attacks' by distributed networks of autonomous robotic systems cooperating at machine speed, such as autonomous swarms of drones with facial recognition.[107]