CCDCOE
NATO COOPERATIVE
CYBER DEFENCE
CENTRE OF EXCELLENCE

Automated/Autonomous Incident Response

Vasileios Anastopoulos, PhD
Davide Giovannelli, LL. M.

NATO CCDCOE

Tallinn 2022

About the authors

This research report is co-authored by Vasileios Anastopoulos and Davide Giovannelli. Chapters 1 to 5 were authored by the former and Chapter 6 by the latter. The conclusions of this work, Chapter 7, were co-authored by both researchers.

CCDCOE

The NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) is a NATO-accredited knowledge hub offering a unique interdisciplinary approach to the most relevant issues in cyber defence. The heart of the Centre is a diverse group of international experts from military, government, academia and industry, currently representing 32 sponsoring and contributing nations.

The CCDCOE maintains its position as an internationally recognised cyber defence hub, a premier subject matter expert and a fundamental resource in the strategic, legal, operational and technical aspects of cyber defence. The Centre offers thought leadership on the cutting edge of all aspects of cyber defence and provides a 360-degree view of the sector. The Centre encourages and supports the process of mainstreaming cybersecurity into NATO and national governance and capability, within its closely connected focus areas of technology, strategy, operations and law.

The Centre is staffed and financed by Austria, Belgium, Bulgaria, Canada, Croatia, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, Montenegro, North Macedonia, the Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, South Korea, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States.

The CCDCOE is home to the Tallinn Manual 2.0, the most comprehensive guide for policy advisers and legal experts on how international law applies to cyber operations carried out between and against states and state actors. Since 2010 the Centre organises Locked Shields, the biggest and most complex technical live-fire challenge in the world that annually enables cyber security experts to enhance their skills in defending national IT systems and critical infrastructure under real-time attacks. The focus is on realistic scenarios, cutting edge technologies and simulating the entire complexity of a massive cyber incident, including strategic decision-making, legal and communication aspects.

The CCDCOE hosts the International Conference on Cyber Conflict, CyCon, a unique annual event in Tallinn, joining key experts and decision-makers of the global cyber defence community. The conference, taking place in Tallinn since 2009, attracts every spring more than 600 participants. The CCDCOE is responsible for identifying and coordinating education and training solutions in the field of cyber defence operations for all NATO bodies across the Alliance.

NATO-accredited centres of excellence are not part of the NATO Command Structure.

www.ccdcoe.org
publications@ccdcoe.org

Disclaimer

This publication is a product of the NATO Cooperative Cyber Defence Centre of Excellence (the Centre). It does not necessarily reflect the policy or the opinion of the Centre or NATO. The Centre may not be held responsible for any loss or harm arising from the use of any information contained in this publication and is not responsible for the content of the external sources, including external websites referenced in this publication.

Table of Contents

Research Methodology

For the compilation of this document design science research methodologies were employed. Literature search patterns [1] were used to research the area of artificial intelligence applications in the field of incident response, organise them, identify gaps and build awareness of developments in industry and practice. A literature base of recent artificial intelligence publications was formed using the Internet, conference proceedings, books and journals.

1. Introduction

Artificial intelligence (AI) has existed for a long time in a way that seems to affect every aspect of our lives in a modern society, but it is only recently that its applications have been made known to the public. AI is already present in many fields including education, agriculture, health and medicine, manufacturing and transportation.

Cyberspace [2] as ‘a global domain within the information environment consisting of the interdependent network of information systems infrastructures including the Internet, telecommunications networks, computer systems, and embedded processors and controllers’, is omnipresent within everyday activities, but its use for malicious acts has also risen the need for cybersecurity [2], ‘the ability to protect or defend the use of cyberspace from cyberattacks’. AI applications are already present in cyberspace, used both by the attackers and the defenders. It could render cyber attacks more successful, leveraging, for example, its ability to replicate natural language and thus making phishing emails more successful, or developing autonomous cyber weapons that could attack and self-replicate. It could also help defenders in detecting anomalies and quickly addressing vulnerabilities and misconfigurations. Currently, there is a controversy about the impact of AI in cyberspace, with one study warning that it could drive to more aggressive and destabilising engagements between nations [3], while another [4] states that attackers will be less likely to employ AI due to its constraints, flaws and limitations unless they see unique benefits. What is clear, though, is that ‘while the discussion regarding autonomy in the physical world is largely about systems that are not quite yet in operation, in the cyber-world, autonomy is already a reality’ [5].

According to recent surveys [6], when an incident response is led by humans it is no longer possible to keep up with the speed, scale and sophistication of automated cyber attacks. The need for more sophisticated technologies is emerging with defenders turning their efforts to guarding against AI-powered attacks and by enabling AI defences. Every day more and more security teams rely on AI to stop threats from escalating even at the early stages of a compromise. Organisations employing AI in cyber security report benefits from its application with increased return on investment (ROI) being one of them [7]. The use of AI in incident response enables security teams to identify, investigate and remediate threats a lot faster, while the effort required is also reduced. Reacting in a timely manner is crucial for cyber defence and reducing the human effort required to respond to security events, facilitates the security teams to focus on the cybersecurity aspects they wish to.

Commercial products have already integrated AI technologies to fight against cyber attacks such as spam mail, ransomware and malware. Vendors continue to integrate AI features into their products while new solutions based on AI are on the rise, such as autonomous response to thwart attacks in progress, automation of the investigations process, protection against phishing attacks, endpoint protection and more.

2. Incident Response

Security attacks are constantly rising in number and evolving in complexity and sophistication, posing the need for effective and efficient incident response (IR). There are various benefits to developing an IR capability in an organisation, including the ability to systematically respond to security incidents, minimising their effect and leveraging lessons learned to better prepare for future incidents. There are various approaches and procedures for IR described in the literature, though for the needs of this document the NIST publication is employed [8].

IR is a complex process and requires thorough planning and the availability of resources. Sample requirements for establishing an IR capability are listed below:

- Create an IR plan and policy.
- Develop procedures for incident handling and reporting.
- Establish communication with external parties.
- Team structure and staff model.
- Establish relationships, and communication, between the IR team and other internal, or external, teams.
- Determine the services to be provided.
- Train and staff the IR team.

Organisations should also seek to reduce the security incidents on their infrastructure and be prepared to handle a variety of threats. IR should be considered and handled as an important procedure for the organisation, properly supported by both management and employees. A feedback loop is necessary to allow the organisation to benefit from the lessons learned, to learn from its experience and to systematically evolve and mature its IR capability.

NIST separates the process of IR into four phases:

- **Preparation**. This typically addresses the establishment and training of an IR team and the acquisition of the necessary tools and resources. The team is preparing for incident handling by establishing communications and acquiring the necessary hardware and software and other relevant resources. Prevention of security incidents is also addressed by conducting risk assessments, advancing host and network security, improving malware prevention and establishing user awareness programmes.

- **Detection and Analysis**. Response strategies are defined depending on the type of security incidents the organisation is facing. Attack vectors, signs of an incident and sources of precursors and indicators are handled during the detection phase, followed by incident analysis, incident documentation, incident prioritisation and notification. It is this phase in which the suspicious actions are detected and analysed to verify whether or not it is an attack, and whether it was successful.
- **Containment, Eradication and Recovery**. A containment strategy is defined in this step to minimise the damage and provide the necessary time for proper remediation and decision-making. Evidence gathering and handling and the identification of the attacking hosts are part of the containment phase. Eradication and recovery follow the containment of a security incident, aiming to eliminate possible remaining components of the security incident and restore the systems to normal operation.
- **Post-Incident Activity**. This phase includes the lessons learned process to allow the IR team to improve and better prepare for new threats. The evidence data collected from the previous phases can be used to improve the organisation's IR capability and other processes. Retaining the evidence data is also addressed in this phase, where the organisation has to establish a policy defining the period for which evidence from security incidents should be retained, either for the benefit of the organisation or due to legal or compliance requirements.

3. Artificial Intelligence

The field of AI has gone through several development cycles and its development and adaptation has greatly accelerated over time. This is largely attributed to the availability of large data sets and the advances in computational performance. These advances have led to increased research on AI with algorithms becoming better and more widely used. Considering the range of AI applications, some advise thinking of AI, not as just as a technology, but rather as an enabler of 'AI-enabled' systems [5].

AI is present in everyday activities with applications in problem-solving, learning, reasoning and planning and in perceiving and acting [9]. Among the techniques used in these applications, machine learning (ML) techniques are already used in cyber security with various products and services incorporating relevant features and capabilities [10]:

- **Support Vector Machine** (SVM) is considered to be among the most successful ML techniques for cyber security, especially for tasks relating to Intrusion Detection Systems (IDS). It can be linear or non-linear and it is demanding in memory and training time.
- **Decision Tree** (DT) is a supervised ML technique that is based on a recursive tree structure composed of a root node, path and leaf node. Each path corresponds to possible values of the parent node, while the leaf node corresponds to the predictive category or classified attribute.

- **K-Nearest Neighbor** (kNN) is an unsupervised learning algorithm based on a distance function that is used to measure the dissimilarity of two data instances. The time it requires for training is less than that for other classifiers, although there is a computation time overhead during the process of classification. The main assumption employed by the classifier is that similar data points in the space will be closer to each other than those that are not similar. It is demanding in storage and computation power, while the classifier is sensitive to noisy data and the selected distance function.
- **Random Forest** (RF) belongs to the category of ensemble learning. This combines multiple classifiers to create a hypothesis of a problem and set up a typical result. It is also used for classification and regression purposes and is typically the collection of prediction results that are generated by multiple decision trees. Intrusion detection and email spam are RF applications often found in the literature. When the problems are non-linear, performance is better and the computational cost is less.
- **Naive Bayes** (NB) is based on Bayes' theorem which decomposes the conditional probability of the problem to be analysed. In cyber security, multiple features are dependent on each other, thus the independence condition does not fit well with various types of attacks. The classifier is fast in detection speed and works well with discrete type attributes. There are three techniques under NB: multinomial, Bernoulli and Gaussian.
- **Artificial Neural Network** (ANN) is a network of artificial neurons or nodes. The nodes' connections are modelled as weights, and each node has an activation function that controls the amplitude of the output. The training requires a sequence of forward- and back-propagation cycles. They require much time for their training and they are considered robust to noise, while extensive data collection improves the accuracy of the model.
- **Recurrent Neural Network** (RNN) is a branch of neural networks and its main purpose is to process time-series data and to analyse data streams. An RNN keeps in memory the information gained from previous experiences and previous states and uses them as input for the next states.
- **Convolutional Neural Network** (CNN) is an extension of the ANN composed of a multi-layer network. It is used extensively for tasks such as image recognition and anomaly detection, among others. In the literature, they are applied to intrusion detection and malicious traffic classification with high accuracy.
- **Deep Belief Network** (DBN) is a branch of deep neural networks following a greedy approach in unsupervised learning. They are supposed to process complex information and recognise complex patterns, mimicking the human brain. Each node is connected with all the previous and following nodes, and the input it receives is based on probabilities.
- **Autoencoders** are unsupervised neural networks that reduce the input dimensions and size of the data by compressing and decomposing them. An autoencoder is usually applied as follows: 1) the encoder is used to learn how to compress the data; 2) the bottleneck layer is used to hold the fully compressed data; 3) using the decoder, the

model learns how to perform data reconstruction; and 4) reconstruction loss gauges how close the output is to the targeted output.

- **Reinforcement Learning** (RL) is an ML subdomain where the algorithm has an input for any wrong prediction; the algorithm has to learn the correct answer after trying several possibilities. RL is usually combined with deep learning when complex problems need to be solved. Its applications in cyber security are mostly focused on host intrusion detection, DDoS protection, cyber-physical systems and phishing emails. In RL, there is an agent directly interacting with the environment that formulates its own learning experiences. A reward function allows the agent to filter out the bad decisions; a penalty may be imposed while rewarding the good ones.

4. Artificial Intelligence Applications on Incident Response

The process of finding software bugs and misconfigurations requires tremendous work from various security professionals trying to find and eliminate vulnerabilities that could be exploited by various threat actors.

To overcome this, on 4 August 2016 Defence Advanced Research Projects Agency (DARPA) hosted the world's first all-machine cyber hacking tournament, the Cyber Grand Challenge [11]. It was a competition for the creation of automatic defence systems that would be capable of reasoning about flaws, creating patches and deploying them in real-time in a network. This competition started with more than a hundred teams of hackers and security researchers and concluded with a final event in which seven teams participated. The Cyber Reasoning System of each team automatically identified software flows and identified affected hosts and the score was based on their ability to protect the hosts, scan the network for vulnerabilities and maintain the correct software functionality. The implementation of such systems that act at machine speed and scale would bring significant benefits and could provide expert-level software security analysis and remediation at enterprise scales and machine speeds.

An intelligence-driven cognitive computing security operations centre (SOC) is introduced in [12] that aims to base its operation on exclusively progressive fully automatic procedures. The authors propose a λ-Architecture Network Flow Forensics Framework (λ-NF3) that aims to be an efficient cybersecurity defence framework against adversarial attacks. Two sophisticated ML algorithms are combined in a hybrid ML framework, addressing issues related to network traffic analysis, malware traffic demystification and identification of encrypted traffic. It follows a reactive cyber security strategy for handling adversarial attacks, combining two opposite classifiers to detect potential threats and discard them, requiring minimum human involvement.

Applications of ML are already present in Intrusion Detection/Prevention Systems (IDPS) [13]. The approaches that are followed by those systems can be divided into two categories:

- Approaches based on AI constructed on supervised training algorithms such kNN, decision trees, MLP and SVM and approaches constructed on unsupervised training algorithms, such as k-means clustering, single linkage clustering and y-algorithm.

- Approaches based on computational intelligence, including artificial immune systems, fuzzy logic, genetic algorithms (GA) and artificial neural networks.

The ML types of classifiers that are commonly used in intrusion detection and prevention systems are [14]:

- **Unique classifiers**: for example, fuzzy logic, GA, self-organising maps, kNN, SVM, and neural networks.

- **Hybrid classifiers**: a combination of several training techniques to improve system performance.

- **Ensemble learning techniques**: employ several algorithms to train a model.

A comparison of commercial network intrusion detection/prevention systems that operate employing various ML techniques is available in [15].

In the literature there are various deep learning approaches used on IDS [16], including:

- Deep neural network;
- Feedforward deep neural network;
- RNN;
- CNN;
- Restricted Boltzmann machine;
- DBN;
- Deep autoencoder;
- Deep migration learning;
- Self-taught learning; and
- Replicator neural network.

The application of data mining algorithms for intrusion detection, mainly for anomaly detection, is detailed in [17]. When anomaly detection is employed, various models of normal behaviour can be built baselining the normal behaviour – the normal use of a resource – thus enabling the identification of deviations. There are three methods commonly used for anomaly detection:

- **Statistical**. The activity of a system is observed and a profile created that represents its behaviour. Usually, one profile is made during the training phase and a second during the detection phase. If the two are different to a specific level, an anomaly is identified.
- **Data mining based.** Particular techniques can be used to unfold changes, associations, patterns and structures in data. Data mining techniques used by IDS include clustering, classification, outlier detection and association rules mining.

- **ML-based.** Frequently used techniques are Bayesian network and Markov models.

ML is also applicable to protecting the endpoints of an organisation's infrastructure. Two trends that are popular while using ML for the detection of malware are:

- The detection of malicious applications using an unknown application set; and
- The detection of malware families using a malicious application set.

There are various products for the protection of endpoint systems that include ML features for malware detection providing enhanced features and capabilities compared to 'traditional' antivirus software solutions. For example, in [18] a new approach for cyber threat detection is presented, exclusively using AI methods for the detection of new malware samples. Another example is [19] where decision trees ensemble are used for malware pre-execution detection on a user's computer, and deep learning for the detection of rare attacks and the detection of post-execution behaviour.

ML is also popular in detecting malware samples, with significant advancements being demonstrated. The available literature includes the application of SVMs, decision trees, NB and deep learning algorithms. A sample list of deep learning algorithms [20] that are used in malware detection are as follows:

- Restricted Boltzmann Machine (RBM)
- Deep Belief Networks (DBN)
- Autoencoder
- Convolution Neural Networks
- Recurrent Neural Networks

ML is also applicable to the triage of security incidents during IR. In [21], a new approach to quick IR triage methods is presented, employing unsupervised learning techniques. The case of web access logs is researched, evaluating various dimensionality reduction methodologies and applying the K-mean algorithm, providing accurate results.

Another application of deep learning to the classification of the events handled by a SOC is demonstrated in [22]. Graphical analysis is used for the identification of a new set of features while the classification of the events is performed using a deep neural network model, yielding encouraging results in terms of classification accuracy.

ML has proved to be valuable in detecting spam. Spam detection is not limited to email spam but has applications to blogs, social media and mobile devices, among others [10]:

- **Spam on emails**. Among the classifiers that have proved to perform well on email spam detection are the decision tree ones (All-Dimensions Tree, Decision Stump and Regression Tree. Bayes Net, SVM and J48 have also been used, while studies comparing their performance have resulted that J48 performed better, while DBNs have also performed well.

- **Spam on Blogs**. Techniques that have been evaluated for spam detection blog entries include Random Forests, Decision Trees, NB, k-NN and SVM and logistic regression.
- **Spam on Twitter**. NB, Decision Trees and Random Forests have been evaluated in literature for the detection of spam tweets, which are tweets that contain malicious code and can result in a security threat.
- **Spam on images**. Detecting image-based spam can be performed by applying techniques available in pattern recognition and computer vision.
- **Spam on videos**. In recent literature, SVM is used to detect spam on videos rendering encouraging results.
- **Spam on mobile devices**. Mobile devices, other than making phone calls, are used for email services, short message services (SMS), access to cloud infrastructures and sharing various file types such as images and videos. Researchers have experimented with various ML techniques to detect spam on these services. Techniques such as Random Forests, logistic regression, k-Means clustering, RNN, SVM and NB have performed well.

In [23], the authors proposed a fully automated cyber defence framework that should require no support from humans to detect and mitigate cyber attacks within a complex infrastructure. The proposed framework architecture connects various cyber sensors with the infrastructure devices, the applications and the actuators, through an AI-powered automated team, aiming to dynamically secure the cyber environment.

5. Artificial Intelligence Challenges

The rapid evolution of the cyber security landscape poses challenges for the ML model and techniques [10].

To apply ML algorithms, a large amount of data is needed and considerable and efficient hardware resources. ML models are usually designed and trained against specific cyber attacks. A model cannot perform well in detecting a variety of attacks, or against evolving cyber attacks. Detecting activities that have not previously been seen can prove challenging and such detection activities are technically substantially different from their precursors [10]. Models are usually trained with past features in a dataset, thus the latest attacks may evade the classifiers resulting in reduced detection rate and false positives.

Another challenge for ML is the data sets that are used for training and evaluating models. Most publicly available data sets are not up-to-date for the latest attacks. Privacy concerns and restrictions hinder the release of data that would be valuable for ML, despite the availability of anonymization techniques. In cyber security, the data typically originates from a great variety and heterogeneity of log sources, and this heterogeneity can prove challenging for ML models.

Attacks on the ML itself must also be considered when designing cyber security measures. In the literature, various types of adversarial attacks aim to fool models by supplying deceptive input. The following are some of the attacks currently available in the literature [10]:

- Fast Gradient Sign Method (FGSM)
- Multi-step Bit Coordinate Ascent (BCAk)
- Multi-step Bit Gradient Ascent (BGAk)
- Generative adversarial networks (GAN)
- Carlini and Wagner attack (C&W)

Defences against such attacks are being addressed by researchers with a few summarised as follows [24] [25]:

- **Defensive distillation** adds flexibility to an algorithm's classification process so that the model becomes less susceptible to exploitation.
- **Feature squeezing** performs smoothing transformations of input features to undo adversarial perturbations.
- In **adversarial training**, classification errors caused by adversarial examples are minimised, injecting inputs to the dataset that contain adversarial perturbations with correct output labels.
- In **gradient masking**, the model's sensitivity to small perturbations in inputs is reduced, computing first-order derivatives of the model.
- **Ensemble methods** improve robustness by training multiple classifiers together.
- **Modifying the training process and input data** can improve the robustness of a deep network by continuously inputting new types of adversarial samples while performing adversarial training. This method, however, requires sufficient expressive power and high-intensity adversarial samples.
- **Modifying network.** The introduction of deep contractive networks, the use of gradient regularisation and biologically inspired solutions have been proposed by researchers.
- **Using an additional network**. Universal perturbations have been proposed against adversarial attacks, adding a separate trained network to the original model.

The National Institute of Standards and Technology has drafted a report [25] on the taxonomy and terminology of adversarial machine learning (AML). It identifies the security challenges for AI and particularly for ML the potential for adversarial exploitation of mode~~l~~ sensitivities to adversely affect the performance of ML classification and regression. AML deals with 'the design of ML algorithms that can resist security challenges, the study of the capabilities of attackers, and the understanding of attack consequences' [25]. They follow a risk-based approach resulting in the AML taxonomy being aligned with three dimensions of AML risk assessment (attacks, defences, and consequences). A summary of the taxonomy is depicted in Figure 1, although the details of each category are omitted for brevity.

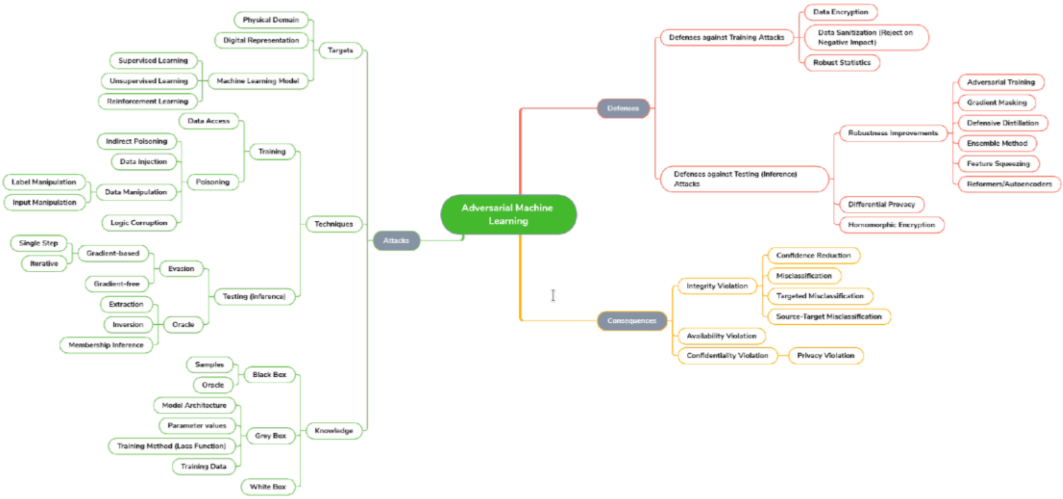

FIGURE 1. TAXONOMY OF ATTACKS, DEFENCES AND CONSEQUENCES IN AML [25]

Considering the attacks that are already present against AI, ongoing research on constructing safe AI systems is available. The researcher in [26] proposed the construction of a privacy-preserving deep learning system under a distributed training system, enabling the collaboration of multiple parties in learning an accurate neural network model without any leak of the input data sets. The innovation introduced with this work is the selective sharing of deep neural network parameters during the training of the modes. Further research on the proposed system resulted in the addition of homomorphic encryption protecting the communication among the components, though increasing the communication overhead [27]. Another approach presented by researchers [28] was a federated learning solution for communication-efficient deep learning networks in decentralised data, followed by the design of a secure aggregations protocol for high-dimensional data in privacy-preserving ML [29].

6. Artificial Intelligence's Legal Implications

AI's legal implications will command significant attention from legislatures and judges around the world over the coming decades. This section will expose the extent of the problem and introduce the trends of existing legislation and case law. Three aspects will be considered: decision-making transparency, civil liability and criminal responsibility. AI's legal implications will not be limited to military systems employing such technology, as AI systems use by private enterprises and public administrations is already raising important legal questions. AI use for military purposes has, however, additional and specific legal implications. Thus, the analysis will start from the general legal framework applicable to all AI systems.

Decision-making transparency

As with any powerful technology, the use of AI systems in our society raises several ethical challenges, for instance relating to their impact on people and society, decision-making capabilities and safety [30] Thus, not surprisingly, the EU Council conclusions of 21 October 2020 emphasised that, with AI, *challenges such as opacity, complexity, bias, a certain degree of unpredictability and partially autonomous behaviour need to be addressed in order to ensure the compatibility of automated systems with fundamental rights and to facilitate the enforcement of legal rules*" [31]. In accordance with EU Council instructions, the European Commission issued a proposal for a Regulation on AI as of April, 2021 (the AI Act) [32]. The proposal is still under evaluation by the Council and the European Parliament, but its analysis could be relevant to understanding the main trends of future legislation on AI. The latter conclusion, moreover, seems to be relevant also for AI systems for military purposes. Although AI Act shall not apply to AI systems developed or used exclusively for military purposes, the provisions of the Act could likely assume as a sort of standard for AI discipline, thereby implying a spill over effect unless a dedicated military regulation will explicitly derogate, in whole or in part, from the AI Act. Additionally, AI military systems should, in principle, follow a discipline equivalent, although not necessarily identical, to that which the AI Act applies to high-risk AI systems because of the public and governmental nature of military activities and the likelihood of affecting the rights of individuals by actions by the armed forces.

On decision-making transparency, the AI Act proposes:

- **Article 13**. High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately.
- **Article 52**. Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use.

These provisions seek to ensure the effectiveness of rights such as the right to an effective remedy and to a fair trial that could be hampered in case of AI systems are not sufficiently transparent, explainable and documented.

The need to ensure decision-making transparency of AI systems has been also recognised within the existing legal framework by several judicial decisions in EU Member States, both with private contracts [33] and administrative decisions [34]. Decision-making transparency is already required for automated individual decision-making unless the data subject's explicit consent has been obtained or the activity is authorised under existing law to which the controller is subject. Decision-making transparency means also that a court could be asked to assess whether an algorithm system is designed in a sufficiently transparent and verifiable manner, as stated by the District Court of The Hague [35].

For a stand-alone high-risk AI system, the Act proposes a new compliance and enforcement system which will generally include internal control checks by providers (with the exception of remote biometric identification systems that would be subject to third party conformity assessment). After the provider has performed this *ex ante* conformity assessment, it should register those systems in an EU database, to increase public transparency and oversight and strengthen the *ex post* supervision of AI systems by competent authorities.

Although the AI Act proposal is not intended to be applicable to AI systems developed or used exclusively for military purposes, it would be appropriate for the development of such systems to follow a conformity assessment capable of ensuring equivalent protection of the one proposed by the Commission particularly given need to comply with Article 36 of the Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977. Under the Protocol, the development, acquisition or adoption of a new weapon or means or method of warfare, a High Contracting Party must determine whether its employment would, in some or all circumstances, be prohibited by any applicable rule of international law applicable.

In the same line of thinking, to achieve a fair balance between all the different interests at stake, the Alliance recently agreed *NATO Artificial Intelligence Strategy* [36] where the following six principles of responsible use of artificial intelligence in defence have been identified:

> *A. Lawfulness: AI applications will be developed and used in accordance with national and international law, including international humanitarian law and human rights law, as applicable.*
>
> *B. Responsibility and Accountability: AI applications will be developed and used with appropriate levels of judgement and care; clear human responsibility shall apply to ensure accountability.*
>
> *C. Explainability and Traceability: AI applications will be appropriately understandable and transparent, including through the use of review methodologies, sources, and procedures. This includes verification, assessment and validation mechanisms at either a NATO and/or national level.*
>
> *D. Reliability: AI applications will have explicit, well-defined use cases. The safety, security, and robustness of such capabilities will be subject to testing and assurance within those use cases across their entire life cycle, including through established NATO and/or national certification procedures.*
>
> *E. Governability: AI applications will be developed and used according to their intended functions and will allow for: appropriate human-machine interaction; the ability to detect and avoid unintended consequences; and the ability to take steps, such as disengagement or deactivation of systems, when such systems demonstrate unintended behaviour.*
>
> *F. Bias Mitigation: Proactive steps will be taken to minimise any unintended bias in the development and use of AI applications and in data sets*.

Civil liability

According to the European Parliament (EP), there is no need for a complete revision of the well-functioning liability regimes, but the complexity, connectivity, opacity, vulnerability, capacity for upgrade, self-learning capability and potential autonomy of AI systems and the multitude of actors involved represent a significant challenge to the effectiveness of Union and national liability framework provisions. It thus considers that specific and coordinated adjustments to the liability regimes are necessary to avoid a situation in which persons who suffer harm or whose property is damaged end up without compensation. EP opinion is based on the conclusion that existing fault-based tort law in Member States generally offers sufficient protection for those that have suffered harm or damage to their property caused by a third party. Concerning civil liability claims against the operator of an AI system, the EP affirms the principle '*that the operator's liability is justified by the fact that he or she is controlling a risk associated with the AI system, comparable to an owner of a car; considers that due to the AI system's complexity and connectivity*' [37]. However, the EP also recognised that *'it seems reasonable to set up a common strict liability regime for those high-risk autonomous AI systems*' [37]. Strict liability is commonly understood as the legal responsibility to compensate third parties for damage or injuries that a product has caused, regardless of intent or mental state were when committing the action. As pointed out by the Expert Group on Liability and New Technologies set up by the European Commission:

> *Only the strict liability of producers for defective products, which constitutes a small part of this kind of liability regimes, is harmonised at EU level by the Product Liability Directive, while all other regimes – apart from some exceptions in specific sectors or under special legislation – are regulated by the Member States themselves* [38].

Thus, at this stage, it is not possible to rule out that an AI Act proposal could be modified during the legislative approval to define a comprehensive strict liability regime for AI systems.

Criminal responsibility

To attribute criminal responsibility, two elements need to be satisfied: *actus reus*, which is the act complained of itself, and *mens rea* which is the internal or mental element, which can be direct or indirect intent, forms: direct intent, indirect intent, recklessness or negligence. As the level of autonomy of AI systems grows, it may be difficult to attribute these two elements to a particular person. It is not surprising, therefore, that some nations have already adopted legislation to regulate criminal responsibility in the use of AI systems such as automated vehicles [39]. Thus, for AI systems, the individual who decides to 'pull the trigger' might no longer be solely responsible for that decision. In any event, with criminal offences arising from the use of AI systems, including autonomous weapons systems, it will be always necessary to identify an individual responsible and so ensure the rule of law is upheld. The Guiding Principles proposed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System [40] includes the following:

1. Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system;
2. Accountability for developing, deploying and using any emerging weapons system in the framework of the CCW (Certain Conventional Weapons) must be ensured in

accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control.

7. Conclusions

For a long time, the IR process has been driven and executed by people. Automation in the execution of cyber attacks has greatly increased the pace with which attacks are now performed, making it difficult for human analysts to follow. Alert fatigue is a common problem amongst security teams that are overwhelmed by the volume and speed of nowadays automated cyber attacks. AI rises as a solution to tackle this problem, being already present in the field of cybersecurity, both in literature and security products.

Among the list of topics that AI encompasses, ML already demonstrates its applications in IR. Common uses include intrusion detection and prevention systems, endpoint protection, network intrusions detection, malware detection, incident triage and of course spam protection, to name a few. AI is also used as an offensive tool for carrying out cyber attacks, leading to the necessity of leveraging AI for defence as a means of tackling the speed and volume of such attacks. It is equally important, though, to consider the AI itself as a target for cyber attack. Techniques for attacking AI systems are already available in the literature and mitigations are being researched.

Consideration must also be given to the need to invest in the latest technologies, to sharpen the technological edge and to maintain NATO’s technological superiority. As authoritatively affirmed by NATO Deputy Secretary General:

> *Because NATO's ability to innovate, is what has guaranteed our military superiority, our technological edge. This is the essential part of deterrence and defence. We have done this brilliantly over the last seven decades. But now our dominance is the political West is being challenged. Because other nations like China or Russia that do not share our same values, the same values like we do, are developing new technologies from hypersonic missiles to autonomous systems to artificial intelligence or cyber warfare. And we risk, if we're not careful, and don't work together, we risk a second Sputnik moment where we suddenly find that we have been outpaced* [41].

AI offers benefits for the IR process, but the ethical dilemmas that sometimes arise from its use and possible gaps in national and international legislation show that we must keep the human analyst on top of any automated or autonomous IR system.

References

[1] Vijay K. Vaishnavi and William Kuechler, Design Science Research Methods and Patterns: Innovating Information and Communication Technology, 2nd ed. CRC Press, Inc. Boca Raton, FL, USA, 2015.

[2] Joint Task Force Transformation Initiative Interagency Working Group, "'NIST SP 800-30, Rev.1 Guide for Conducting Risk Assessments."' Sep. 2012. Accessed: Jun. 06, 2018. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-30/rev-1/final

[3] Wyatt Hoffman, "'AI and the Future of Cyber Competition,"' Center for Security and Emerging Technology, Jan. 2021. https://cset.georgetown.edu/publication/ai-and-the-future-of-cyber-competition/ (accessed Aug. 27, 2021).

[4] Ben Buchanan, John Bansemer, Dakota Cary, Jack Lucas, and Micah Musser, "'Automating Cyber Attacks,"' Center for Security and Emerging Technology, Nov. 2020. https://cset.georgetown.edu/publication/automating-cyber-attacks/ (accessed Aug. 27, 2021).

[5] U. Franke, "'Artificial Intelligence diplomacy | Artificial Intelligence governance as a new external policy tool,"' p. 55.

[6] "'Preparing for AI-enabled cyberattacks."' MIT Technology Review Insights. Accessed: Apr. 08, 2021. [Online]. Available: https://technologyreview.com/

[7] S. B. Atiku, A. U. Aaron, G. K. Job, F. Shittu, and I. Z. Yakubu, "'Survey On The Applications Of Artificial Intelligence In Cyber Security,"' vol. 9, no. 10, p. 6, 2020.

[8] Paul Cichonski, Tom Millar, Tim Grance, and Karen Scarfone, "'NIST SP 800-61, Computer Security Incident Handling Guide, Rev.2-SP800-61.pdf."' Aug. 2012. Accessed: Apr. 19, 2018. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r2.pdf

[9] Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. Pearson, 2015.

[10] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, and M. Xu, "'A Survey on Machine Learning Techniques for Cyber Security in the Last Decade,"' 2020, vol. 8, pp. 222310–222354. doi: 10.1109/ACCESS.2020.3041951.

[11] "'Cyber Grand Challenge."' https://www.darpa.mil/program/cyber-grand-challenge (accessed May 11, 2021).

[12] K. Demertzis, N. Tziritas, P. Kikiras, S. L. Sanchez, and L. Iliadis, "'The Next Generation Cognitive Security Operations Center: Adaptive Analytic Lambda Architecture for Efficient Defense against Adversarial Attacks,"' Jan. 2019, vol. 3, p. 6.

[13] C. Nilă, I. Apostol, and V. Patriciu, "'Machine learning approach to quick incident response,"' 2020, pp. 291–296. doi: 10.1109/COMM48946.2020.9141989.

[14] A. A. Shah, M. S. Hayat, and M. D. Awan, "'Analysis of Machine Learning Techniques for Intrusion Detection System: A Review,"' 2015, vol. 119, pp. 19–29.

[15] "'Vectra vs. Darktrace, ExtraHop, Cisco and Corelight."' https://www.vectra.ai/discover/vendor-comparison (accessed May 13, 2021).

[16] Mohamed Amine Ferrag, Leandros Maglaras, Sotiris Moschoyiannis, and Helge Janicke, "'Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study,"' J. Inf. Secur. Appl., vol. 50, Dec. 2019, doi: http://dx.doi.org/10.1016/j.jisa.2019.102419.

[17] Lidong Wang, "'Big Data in Network Security Systems,"' 2017, vol. 4, pp. 68–74. doi: DOI:10.12691/iteces-4-2-4.

[18] R. Prejbean, "'Cyber Threat Detection using Word Embeddings and Deep Learning."' Bitdefender, Oct. 2019. Accessed: Aug. 27, 2021. [Online]. Available: https://cert.ro/certcon9/presentations/docs/AI/Razvan_Prejbeanu_Cyber_Threat_Detection_using_Word_Embeddings_and_Deep_Learning.pdf

[19] Răzvan Prejbeanu, "'Machine Learning Methods for Malware Detection."' Kaspersky, 2020. [Online]. Available: https://media.kaspersky.com/en/enterprise- security/Kaspersky-Lab-Whitepaper-Machine-Learning.pdf

[20] Ankur Singh Bist, "'Survey of deep learning algorithms for malware detection,"' Int. J. Comput. Sci. Inf. Secur. IJCSIS, vol. 16, Mar. 2018.

[21] C. Nilă and V. Patriciu, "'Taking advantage of unsupervised learning in incident response,"' 2020, pp. 1–6. doi: 10.1109/ECAI50035.2020.9223163.

[22] Nitika Gupta, Issa Traoré, and Paulo Magella Faria de Quinan, "'Automated Event Prioritization for Security Operation Center using Deep Learning,"' presented at the 2019 IEEE International Conference on Big Data, 2019. doi: 10.1109/BigData47090.2019.9006073.

[23] R. Meier, A. Lavrenovs, K. Heinäaro, L. Gambazzi, and V. Lenders, "'Towards an AI-powered Player in Cyber Defence Exercises,"' in 2021 13th International Conference on Cyber Conflict (CyCon), 2021, pp. 309–326. doi: 10.23919/CyCon51939.2021.9467801.

[24] J. Li, "'Cyber security meets artificial intelligence: a survey,"' Front. Inf. Technol. Electron. Eng., vol. 19, no. 12, pp. 1462–1474, Dec. 2018, doi: 10.1631/FITEE.1800573.

[25] Elham Tabassi, Kevin J. Burns, Michael Hadjimichael, Andres D. Molina-Markham, and Julian T. Sexton, "'Draft NISTIR 8269, A Taxonomy and Terminology of 20 Adversarial Machine Learning."' NIST, Oct. 2019. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf

[26] R. Shokri and V. Shmatikov, "'Privacy-Preserving Deep Learning,"' in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, New York, NY, USA, Oct. 2015, pp. 1310–1321. doi: 10.1145/2810103.2813687.

[27] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "'Privacy-Preserving Deep Learning via Additively Homomorphic Encryption,"' IEEE Trans. Inf. Forensics Secur., vol. 13, pp. 1333–1345, 2018.

[28] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "'Communication-Efficient Learning of Deep Networks from Decentralized Data,"' ArXiv160205629 Cs, vol. 54, Feb. 2017, Accessed: Sep. 01, 2021. [Online]. Available: http://arxiv.org/abs/1602.05629

[29] K. Bonawitz et al., "'Practical Secure Aggregation for Privacy-Preserving Machine Learning,"' in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, New York, NY, USA: Association for Computing Machinery, 2017, pp. 1175–1191. Accessed: Sep. 01, 2021. [Online]. Available: https://doi.org/10.1145/3133956.3133982

[30] Report of the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence, set up by the European Commission, 8 April 2019.

[31] Council of the EU - Presidency conclusions - The Charter of Fundamental Rights in the context of Artificial Intelligence and Digital Change (, 21 October 2020 – EU doc. 11481/20).

[32] European Commission Proposal for a Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts - COM/2021/206 final, 21.04.2021.

[33] Amsterdam's District Court, cases Uber drivers v. Uber (transparency requests), Uber drivers v. Uber (deactivation) and Ola drivers v. Ola Cabs (transparency requests), 11 March 2021

[34] Italian Council of State judgement no. 2270 of 8 April 2019 and Italian Council of State judgement no. 881 of 4 February 2020

[35] NJCM et al. v The Dutch State (2020) The Hague District Court ECLI: NL: RBDHA:2020:1878 (SyRI). English translation Available at: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878
In the mentioned judgement, the Dutch Court was ask to assess the legitimate use, by the Dutch government, of the System Risk Indication (SyRI) algorithm system in order toto detect various forms of fraud, including social benefits, allowances, and taxes fraud. The said decision held that SyRI use was not compliant with the article 8 of the European Convention on Human Rights (i.e. right to respect for private and family life). Acknowledged that SyRI was insufficiently transparent and verifiable, "*the court [was] of the opinion that the SyRI legislation contain[ed] insufficient safeguards to protect the right to respect for private life in relation to the risk indicators and the risk model which can be used in a concrete SyRI project. Without insight into the risk indicators and the risk model, or at least without further legal safeguards to compensate for this lack of insight, the SyRI legislation provides insufficient points of reference for the conclusion that by using SyRI the interference with the right to respect for private life is always proportionate and therefore necessary, as required by Article 8 paragraph 2 ECHR, in light of its purpose of combating abuse and fraud*".

[36] Summary of the NATO Artificial Intelligence Strategy, 22 Oct. 2021 - https://www.nato.int/cps/en/natohq/official_texts_187617.htm

[37] European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL))

[38] Expert Group on Liability and New Technologies, Report on Liability for Artificial Intelligence and Other Emerging Digital Technologies (European Commission 2019)

[39] For example, French Law n° 2019-486 of 22 May 2019 places liability of the driver of an automated vehicle for failing to take over control when required. *See* also Council of Europe - European Committee On Crime Problems (Cdpc) - Feasibility Study On A Future Council Of Europe Instrument On Artificial Intelligence And Criminal Law, 4 September 2020).

[40] Report of the 2019 session of the Group of Governmental Experts on Emerging

Technologies in the Area of Lethal Autonomous Weapons Systems (Annex IV), 25 September 2019, doc.no. CCW/GGE.1/2019/3

[41] Remarks by NATO Deputy Secretary General Mircea Geoană at the Road to Warsaw Security Forum 2020 conference, 18 Nov. 2020 - https://www.nato.int/cps/en/natohq/opinions_179612.htm?selectedLocale=en