EUROPEAN UNION AGENCY
FOR CYBERSECURITY

ARTIFICIAL INTELLIGENCE AND CYBERSECURITY RESEARCH

ENISA Research and Innovation Brief

JUNE 2023

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the Agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure and, ultimately, to keep Europe's society and citizens digitally secure. More information about ENISA and its work can be found here: www.enisa.europa.eu.

CONTACT

To contact the authors, please use rit@enisa.europa.eu.
For media enquiries about this paper, please use press@enisa.europa.eu.

EDITORS

Corina Pascu (ENISA), Marco Barros Lourenco (ENISA)

AUTHORS

Dr. Stavros NTALAMPIRAS, University of Milan, I; Dr. Gianluca MISURACA, Co-Founder and VP, Inspiring Futures, ES; Dr. Pierre Rossel, President at Inspiring Futures CH

LEGAL NOTICE

ISBN: 978-92-9204-637-8, DOI: 10.2824/808362

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Artificial Intelligence (AI) is a typical dual-use technology, where malicious actors and innovators are constantly trying to best each other's work. This is a common situation with technologies used to prepare strategic intelligence and support decision making in critical areas. Malicious actors are learning how to make their attacks more efficient by using this technology to find and exploit vulnerabilities in ICT systems.

Taking one step further in clarifying this initial statement: with the help of AI, malicious actors can introduce new capabilities that can prolong or even expand cyber threat practises that have been in existence already for a long time. With AI, these capabilities are gradually becoming automated and harder to detect. This study explores some of these capabilities from a research perspective.

In this study, two dimensions of AI have been considered (categorisation explained in Section 4): (a) ensuring a secure and trustworthy AI and preventing its malicious use ('AI-as-a-crime-service' or 'AI to harm') and (b) the use of AI in cybersecurity ('AI use cases' or 'AI to protect').

The use cases of AI in cybersecurity are numerous and growing. Listing them exhaustively is beyond the scope of this study, as research in this area is constantly evolving. However, we present examples of some of these use cases throughout the report to better explain ongoing research efforts in this technology and explore areas where further research is needed.

The **aim of this study is to identify needs for research on AI for cybersecurity and on securing AI,** as part of ENISA's work in fulfilling its mandate under Article 11 of the Cybersecurity Act[1]. This report is one of the outputs of this task. In it we present the results of the work carried out in 2021[2] and subsequently validated in 2022 and 2023 with stakeholders, experts and community members such as the ENISA AHWG on Artificial Intelligence[3]. ENISA will make its contribution through the identification of five key research needs that will be shared and discussed with stakeholders as proposals for future policy and funding initiatives at the level of the EU and Member States.

No prioritisation of research needs is presented in this report. ENISA conducts its annual prioritisation exercise taking into account the overall status of cybersecurity research and innovation in the EU, policy and funding initiatives for cybersecurity research and innovation in the Union and technical analysis on specific topics and technologies. The priorities for 2022 can be found in the ENISA Research and Innovation Brief Report.

Furthermore, in 2022, ENISA conducted a study reviewing the work of 44 research projects, programmes and initiatives on cybersecurity and AI, which were for the most

[1] https://digital-strategy.ec.europa.eu/en/policies/cybersecurity-act, last accessed January 2023
[2] The considerations in this study are the result of literature review, including of ENISA's prior work on AI, for instance "Securing Machine Learning Algorithms": https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms
[3] Ad-Hoc Working Group on Artificial Intelligence Cybersecurity — ENISA (europa.eu).

part funded by the EU's framework programmes over the period 2014 to 2027. The importance of this inventory relates to the specific role played by AI in the cybersecurity research field, given the continuous and intensifying interplay with other technology families. The fundamental question driving this study was whether investments in cybersecurity R&I on AI have enabled Europe to make progress in this area, especially those backed by EU funds. The findings of this study can also be found in the ENISA Research and Innovation Brief Report 2022.

While we recognise the immense potential in AI for innovation in cybersecurity and the many requirements needed to improve its security, we also acknowledge that there is still much work to be done to fully uncover and describe these requirements. This report is only an initial assessment of where we stand and where we need to look further in these two important facets of this technology.

Furthermore, according to the results of the ENISA study on EU-funded research projects on cybersecurity and AI mentioned earlier, the majority of the projects reviewed focused on machine learning techniques. This can be interpreted in two ways: as a sign that the market for such solutions is particularly appreciative of the potential benefits of ML compared to other fields of AI or that, for some reason, research and development in the other fields of AI is not being adequately considered by public funders despite their recognised potential. In this study, we also highlight the need to further explore the use of ML in cybersecurity but also to investigate other AI concepts.

ENISA has followed the steps outlined in the following list to identify the research needs presented in chapter 7.2 of this report.

- Identification from existing research papers of functions and use cases where AI is being used to support cybersecurity activities, presented in chapter 3.
- Identification from existing research papers of areas where cybersecurity is needed to secure AI, presented in chapter 4.
- Review of AI use cases, presented in chapter 5.
- Analysis of open issues, challenges and gaps, presented in chapter 6.
- Identification of areas where further knowledge is required.

These steps were carried out by experts who contributed to this report mainly through desk research, and the results were validated by members of the R&I community.

ENISA prepares these studies with the aim of using them as a tool to develop advice on cybersecurity R&I and present it to stakeholders. These stakeholders are the main target audience of this report and include members of the wider R&I community (academics, researchers and innovators), industry, the European Commission (EC), the European Cyber Security Competence Centre (ECCC) and the National Coordination Centres (NCCs).

TOP 5 RESEARCH NEEDS FOR AI AND CYBERSECURITY

1

[illegible] and optimise the performance [illegible] AI-based tools and technologies used for cybersecurity

2

Incentivise the development of penetration testing tools based on AI [illegible] exploit security vulnerabilities [illegible] assess attack behaviours

3

Development of standardised framework assessing the preservation of privacy and the confidentiality of information flows as well as of the designed system

4

Development of training for practitioners using real-world scenarios

5

Establishing a observatory for AI and cybersecurity threats

Note: More information on these priorities can be found in Chapter 7 of this report

DEFINITION OF TERMS AND ABBREVIATIONS

The following list describes the terms used in this document.

Artificial Intelligence (AI)	There is no commonly agreed definition of AI[4]. Though a common definition is lacking, a number of commonalities may be observed (cf. JRC[5]) in the definitions analysed that may be considered as the main features of AI: (i) perception of the environment, including consideration of the complexity of the real-world; (ii) information processing (collecting and interpreting inputs (in the form of data); (iii) decision-making (including reasoning and learning): taking actions, performing tasks (including adaptation and reaction to changes in the environment) with a certain level of autonomy; (iv) achievement of specific goals.
Artificial Intelligence systems	AI systems are software (that is developed through machine learning approaches and logic- and knowledge-based approaches[6]). In addition, they can, for a given set of human-defined objectives, generate outputs such as content, predictions and recommendations or decisions influencing the environments with which they interact. AI systems may also possibly include hardware systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge or processing the information derived from this data and deciding the best action(s) to take to achieve a given goal[7] [8].
Artificial neural networks (ANNs)	Artificial neural networks (ANNs), usually simply called neural networks (NNs) or neural nets, are computing systems based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain.
Cyber-physical systems (CPS)	Cyber-physical systems (CPSs) are the integrations of computation, communication and control that achieve the desired performance of physical processes.
Decision Tree (DT)	Decision Tree learning is a form of supervised machine learning.

[4] European Commission. Joint Research Centre. AI watch: defining Artificial Intelligence: towards an operational definition and taxonomy of artificial intelligence. Publications Office, 2020. doi:10.2760/382730. URL https: //data.europa.eu/doi/10.2760/382730. The update to this JRC Technical Report in 2021 https://ai-watch.ec.europa.eu/document/download/e90645f1-662e-470d-9af9-848010260b1f_en provided a qualitative analysis of 37 more AI policy and institutional reports, 23 relevant research publications and 3 market reports, from the beginning of AI in 1955 until 2021.

[5] Idem as 4

[6] Commission proposal for an EU Regulation and Council's General Approach on a Draft AI Act, December 2022, https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf. The initial definition in the Commission's Proposal was narrowed down by the Council to distinguish AI from more classical software systems.

[7] Idem as 4. ETSI defines AI (system) as: 'Artificial intelligence is the ability of a system to handle representations, both explicit and implicit, and procedures to perform tasks that would be considered intelligent if performed by a human'.

[8] The legal definition of AI in the draft EU Regulation is work in progress in the EU Parliament.

Deep Learning (DL)	Deep Learning[9] is part of a broader family of machine learning methods based on artificial neural networks (ANNs[10]).
Ensemble methods	Techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model.
Hidden Markov Model (HMM)	Hidden Markov Model (HMM) is a statistical model which is also used in machine learning. It can be used to describe the evolution of observable events that depend on internal factors that are not directly observable. Hidden Markov models (HMMs) originally emerged in the domain of speech recognition. In recent years, they have attracted growing interest in the area of computer vision as well.
K-means clustering	K-means clustering is one of the simplest and most popular unsupervised machine learning algorithms.
Machine Learning (ML)	Machine learning is a subset of AI which essentially employs advanced statistics in order to construct frameworks with the ability to learn from available data, identify patterns and make predictions without requiring human intervention[11].
Naive Bayes' classifier (NB)	Naive Bayes is a popular supervised machine learning algorithm.
Reinforcement learning (RL)	Reinforcement learning (RL) is an area of machine learning concerned with how intelligent agents take actions in an environment in order to maximise the notion of cumulative reward. Reinforcement learning is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning.
Security-by-design	A concept in software engineering and product design that takes security considerations into account at the early stages of product development.
Supervised ML	Supervised learning is a subcategory of machine learning defined by its use of labelled data sets to train algorithms to classify data or predict outcomes accurately.
Support Vector Machine (SVM)	A Support Vector Machine (SVM) algorithm is a supervised learning algorithm used in the classification of training data sets.
Unsupervised ML	One of the three basic machine learning paradigms, together with reinforcement learning and supervised learning, dealing with the process of inferring underlying hidden patterns from historical data[12].

[9] For example, LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015). Deep Learning. Nature. 521 (7553): 436–444. Bibcode:2015 Nature 521.436L. DOI:10.1038/nature14539
[10] For example, Hardesty, Larry (14 April 2017). Explained: Neural networks. MIT News Office. Retrieved 2 June 2022.
[11] Dipankar Dasgupta, Zahid Akhtar, and Sajib Sen. Machine learning in cybersecurity: a comprehensive survey. The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, page 154851292095127, September 2020. doi:10.1177/1548512920951275. URL https://doi.org/10.1177/1548512920951275
[12] Hinton, Geoffrey; Sejnowski, Terrence (1999). Unsupervised Learning: Foundations of Neural Computation. MIT Press. ISBN 978-0262581684.

KEY AI CONCEPTS AND FEATURES

Machine learning is by far the most popular field in AI. It is used in cybersecurity in a variety of ways. Table 1 below depicts the use of AI methods in cybersecurity functions.

ML involves the development of algorithms and statistical models that allow computer systems to learn from experience and improve without having to be explicitly programmed. In this chapter, we categorise the existing methods of ML into two distinct groups: traditional, and neural network-based tools and methods. This type of categorisation is widely used in the literature to show the advantages and disadvantages of each tool.

There are also other ways to make this categorisation, depending on the use of information (supervised vs unsupervised), scope of application (classification, regression and clustering), depth of architecture (shallow vs deep), etc.

Another school of thought should also be mentioned, namely reinforcement learning (RL), a hybrid approach that aims to learn an environment through an agent based on trial and error.

Table 1: AI methods in cybersecurity functions (source: authors)

Security function/AI	DT	SVM	NB	K-means	HMM	GAS	ANN	CNN	RNN	Encoders	SNN
Intrusion detection	X	X	X	X	X	X	X	X	X		X
Malware detection	X	X	X	X				X	X		
Vulnerability assessment	X										
Spam filtering			X								
Anomaly detection					X					X	
Malware classification						X	X				X
Phishing detection							X				
Traffic analysis								X	X		
Data compression										X	
Feature extraction										X	

1.1 TRADITIONAL ML

Traditional ML-based solutions include DT, SVM and K-means clustering which have been widely used in different cybersecurity tasks such as detection of spam[13],

[13] Saumya Goyal, R. K. Chauhan, and Shabnam Parveen. Spam detection using KNN and decision-tree mechanisms in social networks. In 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), pages 522–526, 2016. doi:10.1109/PDGC.2016.7913250.

intrusions[14] and malware[15], or in the modelling of cyber-physical systems[16]. These will be further described in the next sections of this report.

1.1.1 Decision Trees (DT)

DTs have been extensively used for the detection of spam and intrusions[17] due to their capabilities of identifying rules and patters in network traffic data and system activity. A DT realises a series of rules learnt from the available labelled data, organised in a tree-like structure[18]. Various ML techniques such as DT have been used to **detect cyber-attacks**. Since DTs rely on training data from past incidents and occurrences, most of them fail to detect novel types which are not part of the data set.

The space for possible decision trees is exponentially large, leading to 'greedy approaches'[19] that are often unable to find the best tree. DTs do not take into account interactions between attributes, and each decision boundary includes only a single attribute. Special attention is needed to avoid over-fitting or under-fitting[20] (e.g. pre-pruning, post-pruning, etc.) where most research is focused[21].

Overall, DTs are inexpensive to construct, fast at classifying unknown records, easy to interpret for small-sized trees, robust to noise (especially when methods to avoid overfitting are employed) and can easily handle redundancy.

1.1.2 Support vector machines (SVM)

SVM is a type of machine learning algorithm that can be used for classification or regression analysis. It is one of the most prominent algorithms for cybersecurity applications, as it is suitable for addressing both **anomaly detection and pattern recognition tasks** (spam, malware, and intrusion detection[22]). SVMs are known for their robustness to noise.

[14] S. Krishnaveni, Palani Vigneshwar, S. Kishore, B. Jothi, and S. Sivamohan. Anomaly-based intrusion detection system using support vector machine. In Advances in Intelligent Systems and Computing, pages 723–731. Springer Singapore, 2020. doi:10.1007/978-981-15-0199-9_62. URL https://doi.org/10.1007/978-981-15-0199-9_62

[15] Bassir Pechaz, Majid Vafaie Jahan, and Mehrdad Jalali. Malware detection using hidden Markov model based on Markov blanket feature selection method. In 2015 International Congress on Technology, Communication and Knowledge (ICTCK), pages 558–563, 2015. doi:10.1109/ICTCK.2015.7582729.

[16] Cesare Alippi, Stavros Ntalampiras, and Manuel Roveri. Online model-free sensor fault identification and dictionary learning in cyber-physical systems. In 2016 International Joint Conference on Neural Networks (IJCNN), pages 756–762, 2016. doi:10.1109/IJCNN.2016.7727276

[17] B K Nirupama; M Niranjanamurthy, Network Intrusion Detection using Decision Tree and Random Forest. In 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), DOI: 10.1109/ACCAI53970.2022.9752578. Manish Kumar, M. Hanumanthappa, and T. V. Suresh Kumar. Intrusion detection system using decision tree algorithm. In 2012 IEEE 14th International Conference on Communication Technology, pages 629–634, 2012. DOIi:10.1109/ICCT.2012.6511281.

[18] Víctor H. García, Raúl Monroy, and Maricela Quintana. Web attack detection using ID3. In Professional Practice in Artificial Intelligence, pages 323–332. Springer US, 2006. doi:10.1007/978-0-387-34749-3_34. URL https://doi.org/10.1007/978-0-387-34749-3_34. And Sean T. Miller and Curtis Busby-Earle. Multi-perspective machine learning a classifier ensemble method for intrusion detection. In Proceedings of the 2017 International Conference on Machine Learning and Soft Computing - ICMLSC '17. ACM Press, 2017. doi:10.1145/3036290.3036303. URL https://doi.org/10.1145/3036290 3036303.

[19] Approaches based on heuristics leading to a locally optimal solution.

[20] Overfitting mainly happens when model complexity is higher than the data complexity. it means that model has already captured the common patterns and also it has captured noises too. Underfitting happens when model complexity lower than the data complexity. It means this model is unable to capture even common patterns data (signals).e.g. https://medium.com/geekculture/what-is-overfitting-and-underfitting-in-machine-learning-8907eea8a6c4

[21] Bogumił Kami´nski, Michał Jakubczyk, and Przemysław Szufel. A framework for sensitivity analysis of decision trees. Central European Journal of Operations Research, 26(1):135–159, May 2017. DOI:10.1007/s10100-017-0479-6. URL https://doi.org/10.1007/s10100-017-0479-6 .

[22] Baigaltugs Sanjaa and Erdenebat Chuluun. Malware detection using linear SVM. In Ifost, volume 2, pages 136–138, 2013. doi:10.1109/IFOST.2013.6616872; Min Yang, Xingshu Chen, Yonggang Luo, and Hang Zhang. An android malware detection model based on DT-SVM. Security and Communication Networks, 2020:1–11, December 2020. DOI:10.1155/2020/8841233. URL https://doi.org/10.1155/2020/8841233 ; Kinan Ghanem, Francisco J. Aparicio-Navarro, Konstantinos G. Kyriakopoulos, Sangarapillai Lambotharan, and Jonathon A. Chambers. Support vector machine for network intrusion and cyber-attack detection. In 2017 Sensor Signal Processing for Defence Conference (SSPD), pages 1–5, 2017. DOI: 10.1109/SSPD.2017.8233268.

SVMs are difficult to interpret, which means that it can be difficult to understand how the algorithm arrived at its decision (black box model). In addition, SVMs have limited scalability and depend heavily on the choice of kernel. Other challenges SVMs face include sensitivity to outliers in the data, which can have a significant impact on the location and orientation of the decision boundary, and difficulties in classifying a dataset into multiple classes, where some methods such as one vs one or one vs all can be computationally intensive and time-consuming.

1.1.3 Naive Bayes' classifier (NB)

NB is a versatile and effective ML algorithm that is often used in cybersecurity. It can be used to address classification in cybersecurity tasks by adopting statistical theory, more specifically the Bayes' theorem, to calculate the probability of a class when all features are given as input[23].

The biggest advantage of NB is that it can work with very small data sets. It is one of the most popular algorithms for **spam filtering**[24]**, malware and intrusion detection**. In addition, it is relatively simple to implement and frequently used as a classifier.

NB can operate effectively even in poor data environments. If a data set is not available, one can still use it as a classification algorithm. Moreover, it is a robust method for isolated noise points[25]. However, NB is very prone to overfitting.

1.1.4 K-means clustering (Clustering)

K-means is a popular unsupervised type of ML algorithm used for clustering data points into groups based on similarity. Clustering is considered an important concept to help find a structure or a pattern in a set of unknown data. Clustering algorithms such as K-means are meant to process data and discover clusters (data points that can be grouped) when they are present in a data set. Such clusters can be used to extract useful information and to potentially assist in **identifying intrusions, cyberattacks and malware**[26].

A known limitation of K-means is that it assumes that all clusters have equal sizes and variances[27]. Another limitation is that the algorithm is limited to linear boundaries of data.

1.1.5 Hidden Markov Model (HMM)

HMM works with probability distribution over sequences of observations. HMM is commonly used in statistical pattern recognition where the temporal structure is

[23] Saurabh Mukherjee and Neelam Sharma. Intrusion detection using Naive Bayes classifier with feature reduction. Procedia Technology, 4:119–128, 2012. DOI: 10.1016/j.protcy.2012.05.017. URL https://doi.org/10.1016/j.protcy.2012.05.017

[24] A. Sumithra, A. Ashifa, S. Harini and N. Kumaresan, Probability-based Naïve Bayes Algorithm for Email Spam Classification. In 2022 International Conference on Computer Communication and Informatics (ICCCI), DOI: 10.1109/ICCCI54379.2022.9740792

[25] An 'isolated noise point' has features or values which differ a lot from the majority of the points. Since by definition there are very few such points, their values play a very small role in the conditional probability across all the points

[26] Anjly Chanana, Surjeet Singh, and K.K. Paliwal. Malware detection using ga optimized k-means and hmm. In 2017 International Conference on Computing, Communication and Automation (ICCCA), pages 355–362, 2017. DOI: 10.1109/CCAA.2017.8229842.

[27] https://hackr.io/blog/k-means-clustering, last accesses March 2022.

particularly important[28] for classification. It is a powerful tool for detecting weak signals. Unfortunately, the training data must represent the problem very well and be of high quality in order to optimally decide upon and learn the number of parameters of an HMM. HMM can be used in the cybersecurity domain to assist in several tasks namely in **intrusion detection**[29].

1.1.6 Genetic algorithms (GA)

GA is a heuristic search algorithm used to solve search and optimisation problems. This algorithm is a subset of the evolutionary algorithms[30] used in computation. GA employ the concept of genetics and natural selection to provide solutions to problems[31].

GA-based solutions are typically used in optimisation and search problems. GA-based systems have been used in various cybersecurity applications, including **spam and intrusion detection**[32] [33].

One promising area of research is the use of bio-computation for defence purposes, where techniques for predator avoidance and anti-predator can be adapted to cybersecurity applications[34]. Several approaches based on artificial immune systems for **intruder detection**[35] can be found in the literature.

1.2 NEURAL NETWORKS

1.2.1 Artificial neural Networks (ANNs)

ANNs consist of nodes inspired by the structure of the human brain. By default, they consist of three layers, i.e. the input layer, the hidden layer and the output layer, although additional hidden layers can be added depending on the complexity of the problem. ANNs are often referred to as universal approximators because during the learning process the output is controlled in such a way that the error between the desired and the actual output is minimised[36].

[28] Ahmed Hussen Abdelaziz, Steffen Zeiler, and Dorothea Kolossa. Learning dynamic stream weights for coupled hmm-based audio-visual speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(5):863–876, 2015. DOI:10.1109/TASLP.2015.2409785.
[29] Ye Du, Huiqiang Wang, and Yonggang Pang. HMMs for anomaly intrusion detection. In Computational and Information Science, pages 692–697. Springer Berlin Heidelberg, 2004. DOI:10.1007/978-3-540-30497-5_108. URL https://doi.org/10.1007/978-3-540-30497-5_108
[30] https://www.techtarget.com/whatis/definition/evolutionary-algorithm, last accessed March 2022
[31] Georges R. Harik, Fernando G. Lobo, and Kumara Sastry. Linkage learning via probabilistic modeling in the extended compact genetic algorithm (ECGA). In Scalable Optimization via Probabilistic Modeling, pages 39–61. Springer Berlin Heidelberg, 2006. doi:10.1007/978-3-540-34954-9_3. URL https://doi.org/10.1007/978-3-540-34954-9_3
[32] Anas Arram, Hisham Mousa, and Anzida Zainal. Spam detection using hybrid artificial neural network and genetic algorithms. In 2013 13th International Conference on Intelligent Systems Design and Applications, pages 336–340, 2013. DOI:10.1109/ISDA.2013.6920760. Hossein Gharaee and Hamid Hosseinvand. A new feature selection ids, based on genetic algorithm and SVM. In 2016 8th International Symposium on Telecommunications (IST), pages 139–144, 2016. DOI:10.1109/ISTEL.2016.7881798.
[33] Ying Zhang, Peisong Li, and Xinheng Wang. Intrusion detection for IoT based on improved genetic algorithm and deep belief network. IEEE Access, 7:31711–31722, 2019. DOI:10.1109/ACCESS.2019.2903723.
[34] Siyakha N. Mthunzi, Elhadj Benkhelifa, Tomasz Bosakowski, and Salim Hariri. A bio-inspired approach to cybersecurity. In Machine Learning for Computer and Cyber Security, pages 75–104. CRC Press, February 2019. DOI: 10.1201/9780429504044-4. URL https://doi.org/10.1201/9780429504044-4
[35] Ying Zhang, Peisong Li, and Xinheng Wang. Intrusion detection for IoT based on improved genetic algorithm and deep belief network. IEEE Access, 7:31711–31722, 2019. DOI:10.1109/ACCESS.2019.2903723
[36] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. Nature, 323(6088):533–536, October 1986. DOI:10.1038/323533a0. URL https://doi.org/10.1038/323533a0

ANNs have been used in many areas of cybersecurity such as the **detection of fraud, intrusion, spam and malware**[37]. Overall, multilayer ANNs are prone to overfitting if the network is too large. At the same time, model building can be very time consuming, but testing can be very fast. However, they are sensitive to noise in training data and do not handle missing attributes.

1.2.2 Convolutional Neural Networks (CNNs)

CNNs are types of Neural Networks that are specifically designed for image processing tasks, such as object recognition and classification. CNNs adopt Deep Learning (DL)-based approaches that can efficiently model very large data sets. CNNs use a series of convolutional and pooling layers to extract increasingly abstract features from input images. The convolutional layers apply filters to the input image to identify patterns and features, while the pooling layers down sample the feature maps to reduce the computational complexity of the network. The output of the final layer of the CNN is then fed into a fully connected layer that performs the classification task. Their success followed the huge breakthrough in GPUs with significant data processing capacity. However, they can be computationally intensive as they require graphical processing units (GPUs) to train the models.

In cybersecurity, CNNs have been used for **intrusion detection tasks**[38].

1.2.3 Recurrent Neural Networks (RNNs)

RNNs are a type of neural network that is particularly well-suited for sequential data, such as time series or text data. RNNs are designed to handle inputs of variable length, by processing one element at a time while also maintaining an internal state that summarises the previous inputs. This internal state is passed from one time step to the next, allowing the network to capture dependencies and patterns that exist over time.

RNNs are typically used for **intrusion detection** in the KDD99 data sets (see section 2.4) with high-levels of accuracy[39].

1.2.4 Autoencoders

Autoencoders are a type of unsupervised DNN technique that reduces the dimensionality of the original input space to eliminate noise and irrelevant features.

Autoencoders consist of two parts: an encoder that maps the input data into a lower-dimensional representation, and a decoder that maps the encoded representation back to the original input space. During training, the network learns to minimise the difference between the input data and the reconstructed output, by adjusting the weights of the encoder and decoder.

[37] Preeti Mishra, Vijay Varadharajan, Uday Tupakula, and Emmanuel S. Pilli. A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys Tutorials, 21(1):686–728, 2019. DOI:10.1109/COMST.2018.2847722

[38] Dilara Gümü¸sba¸s, Tulay Yıldırım, Angelo Genovese, and Fabio Scotti. A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems. IEEE Systems Journal, pages 1–15, 2020. DOI:10.1109/JSYST.2020.2992966.

[39] Idem as 38.

Apart from compression applications, autoencoders are effective in detecting anomalies by comparing reconstruction losses between known and new data and are therefore very interesting for cybersecurity applications[40] including the **detection of zero-day attacks**[41].

1.2.5 Siamese Neural Networks (SNN)

SNNs are similarity classifiers that use discriminative features to generalise to unknown categories in a given distribution, e.g. to extract features or distinguish whether two categories belong to the same class, or to categorise data into classes that the model has never 'seen' before. This type of neural network can be used for classification tasks.

The architecture of the SNN is more complicated and additional ML feature extraction mechanisms may need to be added. Compared to conventional neural networks, more time is required for training as a large number of combinations of training samples, necessary for the SNN's learning mechanism, are needed to build an accurate model[42].

Siamese neural networks have many applications in image recognition but also for self-supervised learning (SSL)[43]. SNNs can be effective in quantifying how similar or dissimilar two inputs are at facilitating ML tasks, e.g. classification, anomaly detection, etc.

In cybersecurity, SSNs have been applied to tasks such as **malware detection** and **intrusion detection**, by learning feature representations of the input data that capture the relevant characteristics of malware or anomalous network traffic.

1.2.6 Ensemble methods

ML ensemble methods are techniques that combine multiple machine learning models to improve their accuracy and stability. Ensemble methods are popular because they can improve the accuracy of individual models, reduce overfitting and improve robustness. Even though most of the existing literature utilises systems based on a single ML-based tool, there are several scenarios where ensemble methods have been applied[44].

The reasoning behind using ensemble models is to combine model types that exhibit a promising performance across different cases (e.g. attack types, networks, etc.). Such

[40] Temesguen Messay Kebede, Ouboti Djaneye-Boundjou, Barath Narayanan Narayanan, Anca Ralescu, and David Kapp. Classification of malware programs using autoencoders based on deep learning architecture and its application to the Microsoft malware classification challenge (big 2015) dataset. In 2017 IEEE National Aerospace and Electronics Conference (NAECON), pages 70–75, 2017. DOI:10.1109/NAECON.2017.8268747
[41] Hanan Hindy, Robert Atkinson, Christos Tachtatzis, Jean-Noël Colin, Ethan Bayne, and Xavier Bellekens. Utilising deep learning techniques for effective zero-day attack detection. Electronics, 9(10):1684, October 2020. DOI:10.3390/electronics9101684. URL https://doi.org/10.3390/electronics9101684
[42] https://medium.com/codex/vol-2a-siamese-neural-networks-6df66d33180e, last accessed March 2022.
[43] Attaullah Sahito, Eibe Frank and Bernhard Pfahringer, Semi-supervised Learning Using Siamese Networks, 2019 Springer International Publishing, DOI: 10.1007/978-3-030-35288-2_47
[44] Dipankar Dasgupta, Zahid Akhtar, and Sajib Sen. Machine learning in cybersecurity: a comprehensive survey. The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, September 2020. DOI:10.1177/1548512920951275. URL https://doi.org/10.1177/1548512920951275

methods have been used for several applications including **malware detection**[45], **intrusion detection**[46], etc.

1.3 RELEVANCE OF DEEP LEARNING (DL)-BASED APPROACHES

In recent years enormous amounts of work have been undertaken on designing DL-based solutions to be used in cybersecurity applications including **protection and defence**[47]. DL-based solutions have been able to offer excellent performance which is often superior to traditional ML dealing with large data sets and currently constitute the state-of-the-art in many areas.

However, they come with some important limitations that should be considered during development and implementation. The first is the availability and reliability of data sets, i.e. the need for large data sets containing high quality data[48]. The vast majority of the literature focuses on improving state-of-the art performance, while the reliability of data sets is hardly considered.

Current literature proposes reliability criteria[49] [50] such as: a) attack diversity, b) anonymity, c) available protocols, d) complete capture (with payloads), e) complete interaction, f) complete network configuration, g) complete traffic, h) feature set, i) heterogeneity (all network traffic and system logs), j) correct labelling and k) metadata (full documentation of data collection).

Unfortunately the existing reliability criteria focus on intrusion detection, while similar requirements for other cybersecurity applications are yet to be addressed.

A second important aspect to consider in this specific context is the fact that attackers constantly design new types of attacks bypassing existing security systems. This specific problem falls into the area of learning in non-stationary environments and is usually referred to as concept drift[51].

In addition, the system under study might undergo a shift in its nominal operating conditions (a time-variance), where the nominal model needs updating[52]. Such

[45] Sanjay Kumar, Ari Viinikainen, and Timo Hamalainen. Evaluation of ensemble machine learning methods in mobile threat detection. In 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), pages 261–268, 2017. DOI:10.23919/ICITST.2017.8356396.
[46] Anna Magdalena Kosek and Oliver Gehrke. Ensemble regression model-based anomaly detection for cyber-physical intrusion detection in smart grids. In 2016 IEEE Electrical Power and Energy Conference (EPEC), pages 1–7, 2016. DOI:10.1109/EPEC.2016.7771704.
[47] Dilara Gümü¸sba¸s, Tulay Yıldırım, Angelo Genovese, and Fabio Scotti. A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems. IEEE Systems Journal, pages 1–15, 2020. DOI:10.1109/JSYST.2020.2992966.
[48] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. A survey on deep learning. ACM Computing Surveys, 51(5):1–36, January 2019. DOI:10.1145/3234150. URL https://doi.org/10.1145/3234150
[49] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy. SCITEPRESS - Science and Technology Publications, 2018. DOI:10.5220/0006639801080116. URL https://doi.org/10.5220/0006639801080116
[50] Amirhossein Gharib, Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. An evaluation framework for an intrusion detection dataset. In 2016 International Conference on Information Science and Security (ICISS), pages 1–6, 2016. DOI:10.1109/ICISSEC.2016.7885840.
[51] Gregory Ditzler, Manuel Roveri, Cesare Alippi, and Robi Polikar. Learning in nonstationary environments: A survey. IEEE Computational Intelligence Magazine, 10(4):12–25, 2015. doi:10.1109/MCI.2015.2471196.
[52] Cesare Alippi, Stavros Ntalampiras, and Manuel Roveri. Model-free fault detection and isolation in large-scale cyber-physical systems. IEEE Transactions on Emerging Topics in Computational Intelligence, 1(1):61–71, 2017. DOI:10.1109/TETCI.2016.2641452.

changes need to be detected promptly and identified correctly so that protection mechanisms are able to function reliably.

Therefore, learning in non-stationary environments in cybersecurity remains an open subject and novel techniques able to detect and react appropriately to stationarity changes are required for effective and up-to-date security models.

1.4 COMMONLY-USED CYBERSECURITY DATA SETS

The above-mentioned ML-based tools and methodologies are subject to data availability, i.e. data sets, collections of potentially heterogeneous types of information, attributes or features are necessary for creating such solutions. By analysing the available data and discovering existing patterns, one can gain insights regarding nominal state as well as cyberattacks.

Table 2 presents several widely-used data sets by the R&D community to design ML-based tools and methodologies for cybersecurity applications, such as intrusion detection, malware analysis, botnet traffic modelling or spam filtering. The list provided hereunder is not exhaustive[53], as its aim to present some of the most commonly used data sets and their diverse application scenarios.

Table 2: Widely-used cybersecurity data sets

Data set	Description
KDD Cup 99[54]	This is probably the most widely used data set containing 41 features for anomaly detection. It was designed and made publicly available by the Defence Advanced Research Project Agency (DARPA). It includes full-packet data and four categories of attacks, such as DoS, remote-to-local R2L, user-to-remote (U2R) and probing. It has extensively served approaches to intrusion detection.
DEFCON[55]	This data set includes various attacks to assist intrusion modelling competitions held on a yearly basis.
CTU-13[56]	This includes 13 diverse situations of real-world botnet traffic considering the characteristics of both normal and background traffic.

[53] As new data sets are published at a fast pace, the reader is referred for a comprehensive list of related data sets to: Kamran Shaukat, Suhuai Luo, Vijay Varadharajan, Ibrahim A. Hameed, and Min Xu. A survey on machine learning techniques for cybersecurity in the last decade. IEEE Access, 8:222310–222354, 2020. DOI:10.1109/access.2020.3041951. URL https://doi.org/10.1109/access.2020.3041951 ; Dilara Gümüşbaş, Tulay Yıldırım, Angelo Genovese, and Fabio Scotti. A comprehensive survey of databases and deep learning methods for cybersecurity and intrusion detection systems. IEEE Systems Journal, pages 1–15, 2020. DOI:10.1109/JSYST.2020.2992966; and Iqbal H. Sarker, A. S. M. Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. Cybersecurity data science: an overview from machine learning perspective. Journal of Big Data, 7(1), July 2020. DOI:10.1186/s40537-020-00318-5. URL https://doi.org/10.1186/s40537-020-00318-5 .

[54] R.P. Lippmann, D.J. Fried, I. Graf, J.W. Haines, K.R. Kendall, D. McClung, D. Weber, S.E. Webster, D. Wyschogrod, R.K. Cunningham, and M.A. Zissman. Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation. In Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00, volume 2, pages 12–26 vol.2, 2000. DOI:10.1109/DISCEX.2000.821506.

[55] Ali Shiravi, Hadi Shiravi, Mahbod Tavallaee, and Ali A. Ghorbani. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers & Security, 31(3):357–374, May 2012. DOI:10.1016/j.cose.2011.12.012. URL https://doi.org/10.1016/j.cose.2011.12.012

[56] S. García, M. Grill, J. Stiborek, and A. Zunino. An empirical comparison of botnet detection methods. Computers & Security, 45:100–123, September 2014. DOI:10.1016/j.cose.2014.05.011. URL https://doi.org/10.1016/j.cose.2014.05.011

Spam base[5] [57]	This is a collection of e-mails with several thousand instances facilitating email classification purposes.
SMS Spam Collection[6]	This includes a wide variety of SMS messages labelled as spam or not spam.
CICIDS2017[7]	This consists of traffic data recorded at the Canadian Institute for Cybersecurity and provides full-packet data and raw PCAP files. Interestingly, several types of attacks are considered.
CICAndMal2017[8][58]	This consists or trustworthy and malware applications conveniently organised into four classes, i.e. scareware, SMS malware, ransomware and adware. As such, it is able to facilitate the identification of malicious Android applications.
Android Validation[59]	This consists of data characterising relationships existing between various applications organised into false siblings, siblings, cousins, and step-siblings.
IoT-23 data set[60]	This is a data set containing malicious and benign IoT network traffic.

[57] Tiago A. Almeida, José María G. Hidalgo, and Akebo Yamakami. Contributions to the study of SMS spam filtering. In Proceedings of the 11th ACM symposium on Document engineering - DocEng '11. ACM Press, 2011. DOI:10.1145/2034691.2034742. URL https://doi.org/10.1145/2034691.2034742

[58] Esra Calik Bayazit, Ozgur Koray Sahingoz, and Buket Dogan. Malware detection in android systems with traditional machine learning models: A survey. In 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pages 1–8, 2020. DOI:10.1109/HORA49412.2020.9152840.

[59] Hugo Gonzalez, Natalia Stakhanova, and Ali A. Ghorbani. DroidKin: Lightweight detection of android apps similarity. In Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pages 436–453. Springer International Publishing, 2015. DOI:10.1007/978-3-319-23829-6_30. URL https://doi.org/10.1007/978-3-319-23829-6_30

[60] Sebastian Garcia, Agustin Parmisano and Maria Jose Erquiaga, https://doi.org/10.5281/ZENODO.4743746. Data sets available in https://www.stratosphereips.org/datasets-iot23

AI IN CYBERSECURITY

This section summarises the current state-of-the-art in the main uses of 'traditional', long-standing and newer AI applications (deep learning systems), tools and methods in cybersecurity, looking at both sides of the use of AI in the context of cybersecurity requirements, i.e. malicious and virtuous. Below, is a non-exhaustive list of ways AI can be used in cybersecurity:

a. Cyber-criminals exploiting AI to boost their efficacy;
b. Security mechanisms encompassing AI to detect, identify and mitigate the consequences of compromises;
c. Using AI to exploit vulnerabilities in existing AI and non-AI tools and methodologies, e.g. adversarial attacks[61];
d. Using AI during the design of a system to protect existing AI and non-AI tools and methodologies (protection created during system design).

In the first two cases, AI is used as a tool (an attacker can use AI to design the attack), while in the last two cases AI is the actual target (the attack may target an AI-based system). Even though AI-based defence mechanisms address a wide variety of vulnerabilities, they can themselves be points of attack. Attackers use AI not only to orchestrate various cyber threats, but to attack AI-based defence mechanisms by exploiting existing vulnerabilities. The table below identifies the use of AI methods in cybersecurity functions.

Table 3: AI methods in cybersecurity functions (source: authors)

Security function\AI	DT	SVM	NB	K-means	HMM	GAS	ANN

1.5 EXAMPLES OF USE-CASES

AI-based tools and methodologies can be used to detect and identify cyberattacks and mitigate their consequences. Such tools have the potential to deliver satisfactory performance at low cost and in real time. There is a wide range of safeguard techniques and capabilities that can be enabled by AI[62]. AI-based defence mechanisms are increasingly adopted in the cybersecurity domain, e.g. network and data security, endpoint protection, access reliability, etc[63].

Having described the key concepts and key features in chapter 2, the following sections summarise the kinds of AI (task, technique, method) that are at stake in each particular cybersecurity function or operation, such as prevention of attacks, detection of threats and intrusion, response, and recovery from cyberattacks. In order to do so, we review them using the concepts of prevention, detection, response and recovery.

1.5.1 Prevention

AI can be used to assess vulnerabilities in computer systems and networks. ML algorithms are often used in the analysis of data from multiple sources, such as scanners, security logs and patch management systems, to identify vulnerabilities and prioritise remediation efforts.

Deep learning-based fuzzers[64] are now considered the most promising route for the discovery of vulnerabilities compared to traditional ML[65]. Reinforcement learning can search a computer network for vulnerabilities faster than traditional pen-testing tools.

Table 4: AI applications for the prevention of attacks

Task	Example of AI methods, techniques, approaches
Fuzzers	DL
Pen-testing	Reinforcement learning
Vulnerability assessment	NLP, traditional ML

Source: Authors' adaptation based on Micah and Ashton (2021)

ML can also be beneficial in scoring risk in the network, e.g. to determine the severity of a vulnerability. AI can be used to manage user identities and access to computer systems and applications. ML algorithms can be used to analyse user behaviour and

[62] In particular analyses by Columbus, Louis. n.d. 'Protecting Your Company When Your Privileged Credentials Are for Sale'. Forbes. Accessed 23 August 2021. https://www.forbes.com/sites/louiscolumbus/2018/08/21/protecting-your-company-when-your-privileged-credentials-are-for-sale/; - Dilmegani Cem. 2021. 'Security Analytics: The Ultimate Guide [2021 Update]'. 20 August 2018. https://research.aimultiple.com/security-analytics/ ; Capgemini. 2019. 'Reinventing Cybersecurity with Artifical Intelligence : The New Frontier in Digiotal Security'. AI-in-Cybersecurity_Report_20190711_V06.pdf (capgemini.com) and - Jones, Tim. 2019. IBM Developer 'Take a Look at AI and Security and Explore the Use of Machine Learning Algorithms in Threat Detection and Management. (blog). 19 August 2019. https://developer.ibm.com/articles/ai-and-security /.

[63] Capgemini group. Reinventing cybersecurity with artificial intelligence: A new frontier in digital security. Technical report, Capgemini Research Institute, 01 2021. URL https://www.capgemini.com/research/reinventing-cybersecurity-with-artificial-intelligence/

[64] Examples are the deep learning–based programme NeuFuzz. Microsoft has also studied the use of deep learning for fuzzers, see for instance http://arxiv.org/abs/1711.04596

[65] Several teams in DARPA-sponsored Cyber Grand Challenge competitions attempted to use machine learning to identify software vulnerabilities

identify suspicious activity, such as attempted account takeovers or attempts at unauthorised access.

1.5.2 Detection

Most 'traditional' ML applications fall almost entirely into the detection stage i.e. for spam detection, intrusion detection and malware detection, as well the detection of attacks. A great amount of existing works is focused on spam detection in computer networks. E-mail spams consume relevant resources (e.g. bandwidth, storage, etc.) directly reducing the capacity and efficacy of systems and networks.

Another problem which has been extensively addressed by the research community is the detection of malware and intrusions.

Typically, defence mechanisms are designed to address specific types of attack, such as distributed denial of service (DDoS), probe attacks[66], remote to local attacks (R2L)[67], unauthorised access to local super user (U2R)[68], host-based, network-based, ransomware, etc. A great variety of promising ML-based solutions, including supervised and unsupervised approaches, have been employed to address these specific types of attacks[69] [70]. Moreover, bio-inspired algorithms have been used to address the intrusion detection types of problems[71] [72].

In the area of malware detection[73] [74] [75], ML[76] has been used for selecting relevant features revealing the presence of malware as well as methods for detecting anomalies or abnormalities.

Various ML techniques, such as SVM and DT, have also been used to detect cyberattacks, but most of them fail to detect new types of attacks, i.e. attacks that are not part of the data set used in training. In this case, solutions need to approximate the distribution of the available data so that samples that do not belong to the distribution can be detected. For this purpose, adapted versions of existing traditional (one-class SVM, HMM, etc.) and NN-based (ANN, CNN, etc.) solutions can be used.

[66] In probe attacks the attacker scans the network to gather information on computers in order to identify vulnerabilities.
[67] Remote to local attacks (R2Ls) are known to be launched by attackers to gain unauthorised access to victim machines in networks.
[68] An attack by which an attacker uses a normal account to login into a victim system and tries to gain root/administrator privileges by exploiting some vulnerability.
[69] Kamran Shaukat, Suhuai Luo, Vijay Varadharajan, Ibrahim A. Hameed, and Min Xu. A survey on machine learning techniques for cybersecurity in the last decade. IEEE Access, 8:222310–222354, 2020. doi:10.1109/access.2020.3041951. URL https://doi.org/10.1109/access.2020.3041951
[70] The paper A Survey on Machine Learning Techniques for Cyber Security in the Last Decade https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9277523 is a survey where the performance of various research works is discussed.
[71] Anas Arram, Hisham Mousa, and Anzida Zainal. Spam detection using hybrid artificial neural network and genetic algorithm. In 2013 13th International Conference on Intelligent Systems Design and Applications, pages 336–340, 2013. doi:10.1109/ISDA.2013.6920760
[72] Hossein Gharaee and Hamid Hosseinvand. A new feature selection ids based on genetic algorithm and SVM. In 2016 8th International Symposium on Telecommunications (IST), pages 139–144, 2016. doi:10.1109/ISTEL.2016.7881798.
[73] Hamed HaddadPajouh, Ali Dehghantanha, Raouf Khayami, and Kim-Kwang Raymond Choo. A deep recurrent neural network-based approach for Internet of Things malware threat hunting. Future Generation Computer Systems, 85:88–96, August 2018. doi:10.1016/j.future.2018.03.007. URL https://doi.org/10.1016/j.future.2018.03.007
[74] Temesguen Messay Kebede, Ouboti Djaneye-Boundjou, Barath Narayanan Narayanan, Anca Ralescu, and David Kapp. Classification of malware programs using autoencoders based on deep learning architecture and its application to the Microsoft malware classification challenge (big 2015) dataset. In 2017 IEEE National Aerospace and Electronics Conference (NAECON), pages 70–75, 2017. doi:10.1109/NAECON.2017.8268747
[75] Esra Calik Bayazit, Ozgur Koray Sahingoz, and Buket Dogan. Malware detection in android systems with traditional machine learning models: A survey. In 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pages 1–8, 2020. doi:10.1109/HORA49412.2020.9152840.
[76] See Micah and Ashton (2021) for research exploring the use of ML e.g. HMM and DL techniques.

Furthermore, new data needs to be incorporated into the dictionary for future reference and manual analysis. Table 4 below summarises possible uses of AI techniques for detection of threats and intrusion.

Table 5: AI applications for the detection of threats and intrusion (source: authors' elaboration)

Task	Examples of AI techniques
Spam detection	SVM, DT
Intrusion detection	Supervised and unsupervised approaches, bio-inspired algorithms
Malware detection	Standard ML classifiers, DL
Attack detection	SVM, DT

SECURING AI

This report also examines existing approaches to safer AI, to prevent AI from being used to orchestrate cyberattacks or to prevent attacks on AI-based mechanisms and tools. AI systems themselves may be vulnerable to threats due to their own vulnerabilities or vulnerabilities of other interdependent mechanisms.

1.6 AI SECURITY

Security-by-design is a concept in software engineering that emphasises the importance of integrating security principles in the early stages of the design and development of systems and applications. This includes considering security risks and vulnerabilities at every stage of development, from architecture and design to implementation and testing. The following list contains examples of security-by-design practises that can be applied to AI systems:

- Conducting security risk assessments and threat modelling to identify potential vulnerabilities and attack vectors,
- Using secure coding practices and software development frameworks to minimise the risk of coding errors and vulnerabilities,
- Implementing secure data handling practices to protect sensitive data and prevent data breaches,
- Incorporating security testing and validation into the development process to identify and address security issues early on,
- Ensuring that AI systems are designed to be transparent and explainable, so that their behaviour can be audited and verified.

The concepts of security-by-design that apply specifically to AI systems, include:

- Privacy-by-design: this concept emphasises the importance of incorporating considerations of privacy and data-confidentiality into the design and development of AI systems.
- Explainability-by-design: this concept emphasises the importance of designing AI systems that are transparent and explainable, so that their behaviour can be understood and audited by humans.
- Robustness-by-design: this concept emphasises the importance of designing AI systems that are resilient to attacks and errors, and that can continue to function even in the face of unexpected inputs or disturbances.
- Fairness-by-design: this concept emphasises the importance of designing AI systems that are fair and unbiased, and that do not perpetuate or amplify existing societal biases or discrimination.

1.7 AI-POWERED CYBERATTACKS

As AI technology continues to advance, it is likely that we will see more sophisticated and complex AI-powered cyberattacks in the future. For example, a generative adversarial network (GAN), a class of ML frameworks, can be used to generate 'deep fakes' by swapping or manipulating faces or voices in an image or a video.

AI-based algorithms are also able to prepare persuasive spear-phishing emails[77] targeted at individuals and organisations. AI can also be used to enhance the efficiency and effectiveness of malware[78], by improving its ability to evade detection, adapt to changing environments, target specific vulnerabilities, propagate itself and persist on target systems. AI-driven malware can use reinforcement learning techniques to improve itself and perform even more successful attacks.

Attackers can take advantage of training data to generate a 'back door' in the AI algorithm. Attackers can also use AI to help in deciding which vulnerability is most likely to be worth exploiting. These are just a few examples of AI-powered cyberattacks that already raise substantial concern.

1.8 DEFENDING AI-BASED MECHANISMS

AI systems can be susceptible due to their own vulnerabilities or weak points introduced by other interdependent mechanisms. Attacks against AI-based mechanisms can be organised in the following categories[79] (non-exhaustive list).

- Attacks exploiting existing vulnerabilities in popular open-source software libraries, e.g. pytorch, tensorflow, etc.
- Attacks poisoning training data. Here, it is assumed that the attacker has access to the training data and is able to alter them and introduce manipulations such as wrong labels so that the AI system, trained on poisoned data, carries out processing and/or predictions following the attacker's interests.
- Adversarial attacks, where usually the AI system under attack is a deep neural network. Here, the attacker introduces minor alterations to the test examples in order to alter the prediction of the AI system in a targeted or untargeted manner, i.e. steering the prediction towards a given desired class or to any class other than the correct one.
- Reverse-engineering the trained model based on publicly accessible query interfaces, e.g. model stealing, model inversion and membership inference.

Several approaches have been proposed in literature to secure and protect AI-based mechanisms from such malicious attempts. These approaches include the following.

[77] https://www.wired.com/story/ai-phishing-emails/, last accessed March 2023.
[78] Cong Truong Thanh and Ivan Zelinka. A survey on artificial intelligence in malware as next-generation threats. MENDEL, 25(2):27–34, December 2019. doi:10.13164/mendel.2019.2.027. URL https://doi.org/10.13164/mendel.2019.2.027
[79] ENISA Artificial Intelligence Cybersecurity Challenges, 2020, available at https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges/@@download/fullReport

- Securing the software packages that were used and verifying the validity of the training data[80].
- Approaches addressing adversarial attacks[81] [82] that in general are ad-hoc and focused on a specific type of attack assumed to be known a-priori. This is due to the size of the adversarial attack generation space which is potentially of large dimensions. As such, both traditional and neural network-based ML approaches can be used depending on the specifications of the problem-at-hand.

[80] D. Gümüşbaş, T. Yıldırım, A. Genovese and F. Scotti, A Comprehensive Survey of Databases and Deep Learning Methods for Cybersecurity and Intrusion Detection Systems, in IEEE Systems Journal, vol. 15, no. 2, pp. 1717-1731, June 2021, DOI: 10.1109/JSYST.2020.2992966.
[81] Idem footnote 79
[82] Yunfei Song, Tian Liu, Tongquan Wei, Xiangfeng Wang, Zhe Tao, and Mingsong Chen. Fda3: Federated defense against adversarial attacks for cloud-based IoT applications. IEEE Transactions on Industrial Informatics, ages 1–1, 2020. DOI:10.1109/TII.2020.3005969

SELECTED CASE STUDIES

Four focus areas were examined because of their strong interdependence with AI and cybersecurity, namely the next-generation of telecommunications (6G), cyber biotechnology, the Internet of Things (IoT) and cyber-physical systems (CPS). As some of these areas are still at an early stage of development (at least the first two), there is an expectation that AI will contribute to increasing their potential. This assumption is justified not only in terms of potential but also as regards security.

However, existing cybersecurity tools that use AI may not be adequate in securing these technologies and areas. The use of AI in new contexts needs to be evaluated and often adapted, especially when it learns from data describing attack patterns in new attack surfaces[83]. However, this requires sufficient amounts of reference data to train the models, which may not be yet available due to the novelty of these technologies and domains.

1.9 NEXT GENERATION OF TELCOMMUNICATIONS[84]

In this section we will examine how 5G, beyond-5G and 6G can equally benefit and be at risk from the use of AI. Some promising AI capabilities[85] to support 5G cybersecurity are listed below (not exhaustively):

- optimising resources and dynamic arbitrations, especially in a situation of massive multi-mobility[86],
- improving the management and coordination of algorithms[87],
- improving the 'learning curve' in the management of cybersecurity issues, in particular with the detection of anomalies, e.g. potentially linked to malware, or even attack patterns already listed[88],
- helping to develop more agile and automated capabilities, able to react to subtly changing or threatening situations[89],
- helping to develop security mechanisms by creating trust models, device security and data assurance to provide systematic security for the whole 5G-

[83] Pujolle, Guy (2020). Faut-il avoir peur de la 5G. Paris, Larousse, p. 217-219.
[84] In this chapter, we will leave aside the issue of AI-based facial recognition and surveillance using 5G, a full topic in itself, with growing concerns and technological power.
[85] Haider, Noman; Baig, Muhammad Zeeshan; Imran, Muhammad. 2020. Artificial Intelligence and Machine Learning in 5G Network Security: Opportunities, advantages, and future research trends. arXiv:2007.04490, based upon the 3GPP Technical Specifications Group Services and Systems Aspects.
[86] When numerous mobile agents need to have almost simultaneous access to telecom services, the amount of data to be transferred and monitored has to be supported by AI, as well as the meta surveillance of how this can be subject to attacks, with the forms of attacks themselves being a source of AI learning.
[87] See our comment on Pujolle's explanation above (op. cit.).
[88] In fact, this is how AI can become more and more involved in the defence of 5G hubs and even 5G terminals, i.e. making increasingly better use of past attack (of being attacked) experience.
[89] This reaction time or time management issue is almost by itself such a problem (as attackers also tend to use AI to see how systems defend themselves against attacking probes), that 6G higher expected performance and cybersecurity provisions seem inevitable (see for that Gurtov, Andrei (2020). Network security architecture and cryptographic technologies reaching for post-quantum era., in 6G White Paper: Research Challenges For Trust, Security And Privacy, University of Oulu, Finland, 6G Research Visions, No. 9, 2020, in particular, p. 16, where the author emphasizes the value of AI to provide the dynamicity to match 6G needs for cybersecurity.)

IoT network, involving both classical means and as suggested above, for complex situations and big-data analysis schemes,

- deploying capabilities that reinforce security functions even in the absence of data from actual attacks, e.g. using GAN (generative adversarial networks).

However, AI can pose several challenges to a 5G infrastructure. According to Suomalainen et al. (2020) [90], there are many vulnerabilities, and more research, experimentation and collective learning initiatives are needed to make 5G more secure. Several issues related to the use of AI in 5G were highlighted in this paper, such as the possibility of qualifying the risks of a given situation by its 'explainability'.

Many of the above expectations are unlikely to be fully realised until the next generation of communications (beyond-5G/6G). The development of 6G is expected to reach technological maturity and standardisation towards the end of this decade. At this stage, it is important to remember that this area of research is still far from standardised for cybersecurity functions and specifications. Nevertheless, a key component in the 6G architecture will undoubtedly be the use of AI capabilities, as suggested by the 6G White Paper (Gurtov, op. cit.).

It is expected that 6G will be 'AI-enabled' in the sense that it will rely on AI for its core function, the physical layer, and will enable a wide range of new AI-based applications with the necessary real-time adaptability and will also be made more secure against opportunistic AI-based attacks. Key areas of the 6G architecture will rely on AI to some (high) degree, e.g. an intelligent real-time edge for enhanced real-time control capability at scale, distributed AI for decentralised decision-making, intelligent radio frequency allocation for dynamic configuration of radio frames, intelligent network management for end-to-end automation of network management[91]. Some examples of new AI-based capabilities are multisensory augmented reality (XR), connected robotics and autonomous systems (CRAS) or wireless brain-computer interaction (BCI) [92].

1.10 INTERNET OF THINGS (IOT) AND INTERNET OF EVERYTHING (IOE)[93]

In the context of IoT, the aspects of complexity, speed and efficiency are promoted by AI. The next generation of IoT will most likely be driven by industry needs. To provide just one example, AI can help improve security measures by checking for intrusions and anomalies and predicting the risk of service outages. For another example, AI plays an important role[94] in the analysis of incoming data and network-wide analytics.

[90] Suomalainen, J., Juhola, A., Shahabuddin, S., Mämmelä, A., & Ahmad, I. 2020. Machine Learning Threatens 5G Security. IEEE Access, 8, 190822 - 190842. https://doi.org/10.1109/ACCESS.2020.3031966
[91] Wang, et al. 2020. 'Security and Privacy in 6G Networks: New Areas and New Challenges'. Digital Communications and Networks 6 (3): 281–91. https://doi.org/10.1016/j.dcan.2020.07.003
[92] Siriwardhana et al. 2021. 'AI and 6G Security: Opportunities and Challenges'. https://doi.org/10.1109/EuCNC/6GSummit51104.2021.9482503
[93] IoT systems are known to convey a series of vulnerabilities, but AI in this domain is just one of the possible tools to detect anomalies or learn from the experience of past attacks and does not differ significantly from its most general use. However, in the current situation, as IoT accounts for a significant amount of cybersecurity incidents, it is important to understand the nature of its vulnerabilities, regardless of the fact that some of these problems may be mitigated thanks to AI or not.
[94]

AI is undoubtedly an excellent set of tools for mitigating IoT risks, whether by investigating vulnerabilities, anticipating problems (or even predicting them through self-reporting capabilities), controlling cross-network issues, orchestrating traffic flows, and generally reducing risk[95].

A particular source of vulnerabilities lies in the so-called Internet of Everything (IoE), an evolution of the Internet of Things (IoT), which through the architecture of 'tiny cells' becomes a comprehensive ecosystem connecting billions of different devices[96]. These devices are a prime target for attackers[97]. In addition, this architectural feature also poses privacy risks in terms of the collection of location and identity data. Smaller, dense and constantly connected local networks will potentially include body-worn networks, drones and environmental sensors with low levels of security that collect and share highly sensitive information, as we will see for the IoT in general, a problem that the 6G community will have to deal with efficiently.

Security issues arise when Edge Intelligence (Cloud at the Edge) ML deploys tools that are vulnerable to poisoning attacks or other forms of intrusion during their learning process. Intentionally injecting false data or manipulating the logic of the data can lead to errors in interpretation or nefarious behaviour[98]. One theoretical countermeasure to this threat is defence systems that are able to mimic and outperform the attacker.

1.11 CYBERSECURITY IN CYBER-PHYSICAL SYSTEMS (CPS)

Cyber-physical systems are a crucial element in complex technical systems such as power supply systems, water supply networks, transport systems, robotic systems, smart buildings, etc., improving the overall utilisation and control of their components. However, their presence has opened the door for cyberattacks[99] [100]. The purpose of such malicious acts can vary and usually involves the theft, corruption or even destruction of information and/or system components[101]. Needless to say, as we write these lines, there is a real situation with the war in Ukraine that permanently includes these threats to critical infrastructure[102].

There are three methods for detecting cyberattacks in CPSs:

[95] For that, see in particular Hodo, Elike, et al. 2016. 'Threat Analysis of IoT Networks Using Artificial Neural Network Intrusion Detec-tion System'. 2016 International Symposium on Networks, Computers and Communications (ISNCC), May, 1–6. https://doi.org/10.1109/ISNCC.2016.7746067

[96] Presentation at Black Hat USA 2022 of the API ecosystem connecting IoT/IOE devices with functionalities https://i.blackhat.com/USA-22/Wednesday/US-22-Shaik-Attacks-From-a-New-Front-Door-in-4G-5G-Mobile-Networks.pdf

[97] Idem as **Error! Bookmark not defined.**

[98] Benzaid and T. Taleb. 2020. AI for Beyond 5G Networks: A Cyber- Security Defense or Offense Enabler? IEEE Network, Vol. 34, No. 6, Pp. 140–147, 2020. https://doi.org/10.1109/MNET.011.2000088

[99] Sridhar Adepu, Venkata Reddy Palleti, Gyanendra Mishra, and Aditya Mathur. Investigation of cyber-attacks on a water distribution system. In Lecture Notes in Computer Science, pages 274–291. Springer International Publishing, 2020. DOI:10.1007/978-3-030-61638-0_16. URL https://doi.org/10.1007/978-3-030-61638-0_16

[100] Peter Eder-Neuhauser, Tanja Zseby, Joachim Fabini, and Gernot Vormayr. Cyber-attack models for smart grid environments. Sustainable Energy, Grids and Networks, 12:10–29, December 2017. DOI:10.1016/j.segan.2017.08.002. URL https://doi.org/10.1016/j.segan.2017.08.002

[101] Antonello Monti and Ferdinanda Ponci. Electric power systems. In Intelligent Monitoring, Control, and Security of Critical Infrastructure Systems, pages 31–65. Springer Berlin Heidelberg, September 2014. DOI:10.1007/978-3-662-44160-2_2. URL https://doi.org/10.1007/978-3-662-44160-2_2

[102] Just to give an idea, according to the New York Times of Nov. 17, 2022, there had already been, until then, 126 cyberattacks on the Ukrainian power system from the Russians. Russian Attacks on Ukraine's Power Grid Endanger Nuclear Plants, U.N. Agency Says - The New York Times (nytimes.com)

a) signature-based i.e. searching for known patterns of malicious activity in the data stream using a predefined dictionary of attacks[103],

b) anomaly-based i.e. estimating characteristic features of normal behaviour and subsequently detecting deviations that may appear during an intrusion[104],

c) countermeasure-based i.e. adapting the signals involved (by adding information which demonstrates authenticity) so that the task of intrusion detection is simplified[105].

The above methods can be used as a first line of defence if the computational cost is relatively low.

Anomaly-based methods and suspicious correlations with big data should be able to address more complicated cases of malicious events and sophisticated attacks and are considered promising, i.e. for CPS families such as smart grids, vehicular, industrial and medical CPS, and are being explored in the literature[106]). This has more to do with the idea of acceptable confidence in a given system at a given time and context rather than goals for measuring absolute effectiveness[107].

Various ML techniques can be used for modelling and anomaly detection, including NNs, rule-based schemes, predefined suspicious big data filtering schemes[108]. There has been a remarkable increase in research in ML-based solutions due to the widespread development and application of DL/RL algorithms. However, despite these continuous improvements, it seems that the current state of security algorithms cannot quite keep up with the development of novel attacks. This is partly due to the ingenuity of attackers, but also due to the difficulty of defending complex systems that involve not only infrastructures but also all the people inside and outside them, making them true information ecosystems.

1.12 CYBER BIOSECURITY

The increasing convergence of biotechnology and AI is an emerging field for exploitation. An initial attempt to problematise the research area at the intersection of cybersecurity, cyber-physical security and biosecurity resulted in the proposed definition of cyber biosecurity as ‘understanding the vulnerability to unwanted surveillance, intrusions, and malicious and harmful activities, that may occur in or at the interfaces of interconnected life and medical sciences, cyber, cyber-physical, supply chain and infrastructure systems, and the development and implementation of

[103] Hu Zhengbing, Li Zhitang, and Wu Junqi. A novel network intrusion detection system (NIDS) based on signatures search of data mining. In First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008), pages 10–16, 2008. doi:10.1109/WKDD.2008.48.
[104] Jan Neuzil, Ondrej Kreibich, and Radislav Smid. A distributed fault detection system based on IWSN for machine condition monitoring. IEEE Transactions on Industrial Informatics, 10(2):1118–1123, 2014. DOI:10.1109/TII.2013.2290432.
[105] Yilin Mo, Rohan Chabukswar, and Bruno Sinopoli. Detecting integrity attacks on SCADA systems. IEEE Transactions on Control Systems Technology, 22(4):1396–1407, 2014. DOI:10.1109/TCST.2013.2280899.
[106] Felix O. Olowononi, Danda B Rawat, and Chunmei Liu. Resilient machine learning for networked cyber physical systems: A survey for machine learning security to securing machine learning for CPS. IEEE Communications Surveys & Tutorials, 23(1):524–552, 2021. ISSN 2373-745X. DOI:10.1109/comst.2020.3036778. URL http://dx.doi.org/10.1109/COMST.2020.3036778
[107] See in particular for that Siau Keng and Wang Weiyu (2018). Building Trust in Artificial Intelligence, Machine Learning, and Robotics. CUTTER Business Technology Journal (2) (PDF) Building Trust in Artificial Intelligence, Machine Learning, and Robotics (researchgate.net)
[108] Siddharth Sridhar and Manimaran Govindarasu. Model-based attack detection and mitigation for automatic generation control. IEEE Transactions on Smart Grid, 5(2):580–591, 2014. DOI:10.1109/TSG.2014.2298195

measures to prevent, protect against, mitigate, investigate and attribute such threats to security, competitiveness and resilience'[109].

The most important mechanism introduced using AI in biotechnology[110] is the ability to automate complex tasks without direct supervision or to use cyberattacks to exploit bio automation for malicious purposes.

At the same time, they are examples of 'dual-use research of concern' (DURC), i.e. technologies that clearly have positive impacts while opening up new opportunities that can also be exploited for malicious purposes (Pauwels, 2021). A major concern with AI, as discussed earlier, is explainability and the production of replicable and usable knowledge (Jordan et al., 2020). However, it has yet to be demonstrated with real evidence that bio-evolution can pose new specific threats that are not just an extension of the existing potential attack surface. Biometric systems show that it is more about proliferation of cybersecurity deployments than a real paradigm shift, but of course it seems a bit early to close this debate.

[109] Peccoud, J., Gallegos, J. E., Murch, R., Buchholz, W. G., Raman, S. 2018. Cyberbiosecurity: From Naive Trust to Risk Awareness. Trends in Biotechnology, 36(1), 4-7. https://doi.org/10.1016/j.tibtech.2017.10.012

[110] This one is, in fact, a very diversified landscape, with obvious cross-fertilising possibilities among domains of application and disciplines, as, for instance, one of the primary uses of AI in the biofield is assistance to identify and model new proteins of high-potential pharma-oriented molecules.

- how to achieve end-to-end protection (data is particularly at risk when it is in transit[113]);
- how to achieve optimal accuracy under real-world conditions and not in a simulated environment[114];
- the need for computational complexity and 'low-latency operation' to be addressed especially when the system being monitored is of critical importance[115];
- the need to investigate whether the inferred models are valid or biased, or whether there are perceive changes in the time variance[116];
- Ensuring that the security of the protection mechanism is assessed following a standardised framework considering diverse malicious attempts, cases, figures of merit, etc. (security-by-design)[117];
- preservation of privacy e.g. training data and confidentiality of the information flowing in the system so that the characteristics of the system are not exposed indirectly and potentially classified information is not also revealed[118].

1.14 RESEARCH GAPS

The following non-exhaustive list provides the research gaps that were identified in our study:

- Construction of effective AI models with a relatively small amount of data by moving from big data to a small data environment;
- Elaboration on raw data targeting end-to-end solutions where feature engineering and the need for domain expertise (knowledge) is minimised or even eliminated;
- Incorporation of change detection and adaptation mechanisms to address non-stationarities (changes in the time variance of system states);
- Periodical assessment of the validity of the developed model(s) so as to promptly detect and address potential bias(es) which introduce additional vulnerabilities;
- Development of approaches to remove existing biases, imbalances, etc. which may degrade the performance of the model;
- Development of standardised data sets following these requirements in order to reliably reproduce and compare existing AI-based solutions;

[113] Trantidou, et al, 2022, SENTINEL - Approachable, tailor-made cybersecurity and data protection for small enterprises, in PROCEEDINGS 2022 IEEE International Conference on Cyber Security and Resilience (CSR), DOI: 10.1109/CSR54599.2022.9850297.
[114] Kavak et al, 2021, Simulation for cybersecurity: state of the art and future directions, DOI: 10.1093/cybsec/tyab005, Oxford University Press (OUP), Journal of Cybersecurity.
[115] Zhenyu Guan, Liangxu Bian, Tao Shang, and Jianwei Liu. When machine learning meets security issues: A survey. In 2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR), pages 158–165, 2018. doi:10.1109/IISR.2018.8535799. Liu et al, 2022, Complexity Measures for IoT Network Traffic, IEEE Internet of Things Journal, DOI: 10.1109/JIOT.2022.3197323.
[116] Ntalampiras and Potamitis, 2022, Few-shot learning for modelling cyber physical systems in non-stationary environments, DOI: 10.1007/s00521-022-07903-0. Springer Science and Business Media (LLC), Journal Neural Computing and Applications.
[117] Karie et al, 2021 A Review of Security Standards and Frameworks for IoT-Based Smart Environments, DOI: 10.1109/ACCESS.2021.3109886, IEEE
[118] Domingo Ferrer and Alberto Blanco-Justicia, 2020, Privacy-Preserving Technologies, DOI: 10.1007/978-3-030-29053-5_14, Springer International Publishing, The International Library of Ethics, Law and Technology.

- Development of approaches to distinguish malicious attacks from faulty states[119];
- On understanding how the efficacy of AI-based tools and methodologies is altered in terms of both accuracy and computational complexity due to an increase in the scale of the system[120], and consequently an increase in the impact of a cyberattack;
- Modelling interdependent cyber-physical systems in order to assess the impact of vulnerabilities;
- The need for a standardised performance evaluation framework to enable reliable comparison between solutions addressing the same or similar problems;
- Provision of context awareness[121] in ML in order to boost resiliency;
- Bringing 'humans into the loop' e.g. training practitioners using real-world scenarios.

While these research gaps cover AI in general, they are particularly important for cybersecurity applications.

1.15 RESEARCH NEEDS

The following list presents the needs for further research on the use of AI or ML concepts in cybersecurity:

1. test beds to study and optimise the performance of ML-based tools and technologies used for cybersecurity,
2. development of penetration testing tools based on AI and ML to find and exploit security vulnerabilities to assess the behaviour of attackers,
3. development of standardised frameworks assessing the preservation of privacy and the confidentiality of information flows as well as the designed system,
4. development of AI training models for practitioners using real-world scenarios,
5. establishing an observatory for AI and cybersecurity threats.

The tables below present ENISA's proposals for future funding calls based on the needs identified in the list above.

Test-beds to optimise the performance of AI/ML-based tools and technologies used for cybersecurity
Type: AI for cybersecurity

[119] Yannis Soupionis, Stavros Ntalampiras, and Georgios Giannopoulos. Faults and cyber-attacks detection in critical infrastructures. In Critical Information Infrastructures Security, pages 283–289. Springer International Publishing, 2016. DOI:10.1007/978-3-319-31664-2_29. URL https://doi.org/10.1007/978-3-319-31664-2_29 .

[120] Cesare Alippi, Stavros Ntalampiras, and Manuel Roveri. Model-free fault detection and isolation in large-scale cyber-physical systems. IEEE Transactions on Emerging Topics in Computational Intelligence, 1(1):61–71, 2017. DOI:10.1109/TETCI.2016.2641452

[121] Context awareness refers to the ability of the protection mechanism to collect information from its surrounding and interconnected environment in order to adapt to potential changes and incorporate them into its operation. As such, protection quality could be boosted since previously unavailable information would be employed to learn the system model on-the-fly.

Description:

Testbeds are required to study and optimise the performance of ML-based tools and technologies used for cybersecurity.

Objectives:

1. Develop test beds to optimise the performance of AI/ML used in cybersecurity.

Entities:

- Security researchers
- Application developers

Beneficiaries:

- Providers of AI tools and solutions

Existing research:

Much of the existing research effort is focused on achieving optimal accuracy in a simulated environment, which usually does not reflect the performance achieved under real-world conditions. Computational complexity and real-time operation must be taken into account, especially when the system being monitored is of critical importance. In this direction, efforts are needed to improve and construct test environments to study and optimise the performance of ML-based cybersecurity tools and technologies.

Standardised frameworks assessing the preservation of privacy and confidentiality

Type: AI for Security

Description:

Standardised frameworks assessing the preservation of privacy and confidentiality of the information flows as well as of the designed solutions need to be developed.

Objectives:

1. Preservation of privacy, e.g. training data and confidentiality of the information flows in systems so that the characteristics of the systems are not indirectly exposed and potentially classified information is not revealed;
2. Privacy preservation and confidentiality offered by the designed solutions.

Entities:

- Security researchers
- Application developers
- GDPR-related specialists

Beneficiaries:

- Software industry

Existing research:

The preservation of privacy and confidentiality of the information flows and of the designed solutions are issues that are rarely considered.

AI/ML-based penetration testing

Type: AI for cybersecurity

Description:

AI-powered penetration testing

Objectives:

1. Using AI/ML to test a system to find security vulnerabilities that an attacker could exploit and then trying to figure out what an attacker will do.

Entities:

- Security researchers
- Application developers

Beneficiaries:

- Cybersecurity practitioners
- Cybersecurity industry

Existing research:

Threat actors can take advantage of training data by generating a backdoor. They can use AI to find the most likely vulnerability to exploit. Penetration testing can lead to finding vulnerabilities that give outsiders access to the data training models.

There are many automated tools that complement penetration testing tools. These automated solutions have some basic AI capabilities, and these capabilities are gradually increasing thanks to ongoing research and open competitions. For example, the 2016 Cyber Grand Challenge - a DARPA-sponsored competition - challenged people to build hacking bots and compete against each other. These artificially intelligent bots perform penetration tests to look for security vulnerabilities and close them before competing teams can exploit them. For example, Mayhem was able to find, fix and search for intrusions on its host system, while simultaneously finding and exploiting vulnerabilities on rival systems.

As we write this study, Generative Pre-trained Transformer software is emerging first through OpenChat GPT and then with the promises of a handful of competitors. Research

Sharing real-time information on AI and cybersecurity threats, at software and hardware levels, as well as attackers' modus operandi is a must for Europe to function as a coherent defence arena.
Objectives: Develop an inventory of trends and threats at software and hardware levels as well as the modus operandi of attackers.
Entities: • European Cybersecurity Competence Centre
Beneficiaries: • Cybersecurity community
Existing research: Developing an observatory of threats would require developing a network of observatories across the EU and linked to like-minded countries and key partners and organisations. The European Cybersecurity Centre could be an organisation to play such role, provided that this particular 'observation-and-sharing' objective be specified.

CONCLUSIONS AND NEXT STEPS

AI is gaining attention in most quadrants of society and the economy, as it can impact people's daily lives and plays a key role in the ongoing digital transformation through its automated decision-making capabilities. AI is also seen as an important enabler of cybersecurity innovation for two main reasons: its ability to detect and respond to cyber threats and the need to secure AI-based applications.

The EU has long considered AI as a technology of strategic importance and refers to it in various policy and strategy documents. ENISA is contributing to these EU efforts with technical studies on cybersecurity and AI. For example, the cyber threat landscape for AI[123] raised awareness on the opportunities and challenges of this technology. The Agency has already published two studies on this topic and this report will be the third publication aiming to provide a research and innovation perspective of cybersecurity and AI. In preparing these studies, the Agency is supported by the R&I community and has established an *ad-hoc* working group[124] with experts and stakeholders from different fields and domains.

This study makes recommendations to address some of the challenges through research and identifies key areas to guide stakeholders driving cybersecurity research and development on AI and cybersecurity. These recommendations constitute ENISA's advice, in particular to the EC and ECCC, using its prerogative as an observer on the Governing Board and advisor to the Centre. The findings were used to produce an assessment of the current state of cybersecurity research and innovation in the EU and contribute to the analysis of research and innovation priorities for 2022, presented in a separate report.

In this context and as next steps, ENISA will:

1. present and discuss the research and innovation priorities identified in 2022 with members of the ECCC Governing Board and NCCs;
2. develop a roadmap and establish an observatory for cybersecurity R&I where AI is a key technology; and
3. continue identifying R&I needs and priorities as part of ENISA's mandate (Article 11 of the CSA).

[123] ENISA. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges, last access March 2023.
[124] ENISA. https://www.enisa.europa.eu/topics/iot-and-smart-infrastructures/artificial_intelligence/ad-hoc-working-group/adhoc_wg_calls, last accessed March 2023.

TP-07-23-202-EN-N

ABOUT ENISA

The European Union Agency for Cybersecurity, ENISA, is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe. Established in 2004 and strengthened by the EU Cybersecurity Act, the European Union Agency for Cybersecurity contributes to EU cyber policy, enhances the trustworthiness of ICT products, services and processes with cybersecurity certification schemes, cooperates with Member States and EU bodies and helps Europe prepare for the cyber challenges of tomorrow. Through knowledge sharing, capacity building and awareness raising, the Agency works together with its key stakeholders to strengthen trust in the connected economy, to boost resilience of the Union's infrastructure and, ultimately, to keep Europe's society and citizens digitally secure. More information about ENISA and its work can be found at: www.enisa.europa.eu.

ENISA
European Union Agency for Cybersecurity

Athens Office
Agamemnonos 14, Chalandri 15231, Attiki, Greece

Heraklion Office
95 Nikolaou Plastira
700 13 Vassilika Vouton, Heraklion, Greece

enisa.europa.eu

ISBN 978-92-9204-637-8
doi: 10.2824/808362