NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Intelligent Systems for Cyber Defence

An Architecture Framework for Cyber Defence using Artificial Intelligence

Paulo César Prata Oliveira Trilho

Dissertation presented as partial requirement for obtaining the master's degree in Information Management, specialization in Information Systems and Technologies Management.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INTELLIGENT SYSTEMS FOR CYBER DEFENCE

AN ARCHITECTURE FRAMEWORK FOR CYBER DEFENCE USING ARTIFICIAL INTELLIGENCE

by

Paulo César Prata Oliveira Trilho

Dissertation presented as a partial requirement to obtain the master's degree in Information Management with a specialization in Information Systems and Technologies Management

Advisor: Professor Doutor Vitor Manuel Pereira Duarte dos Santos

May 2022

ACKNOWLEDGEMENTS

Firstly, I would like to thank my wife Susana for her support and belief in me when I decided to take on the challenge of signing up for a university course at this stage of our lives. Knowing the huge impact, it would take on our personal lives, on our family time, especially with regards the less attention given to our daughter, it meant an extra load on her throughout the duration of the course and the execution of this thesis. When things got hard, and I was feeling that probably it wasn't possible to have success in the studies while having a very demanding day job, a wife, and a little daughter, when I was thinking of giving up, it was her motivation and belief that made me continue and in the end be successful in this endeavor.

An appreciation word for all the experts who accepted to spare some of their precious time to be interviewed by me and provide their valuable feedback on my work. Professor Marco Reis whom I only know for being my teacher in the cybersecurity course of the post-graduation, was amazingly helpful and supportive in the discussion phase of my work. Julian Williams who works in Vodafone Group cyber security organization was also crucial in his provided feedback and technical expertise. My old ISEL colleague Nuno Pires who I haven't seen for possibly 10 years and is now a cybersecurity expert in CNCS was also fantastic by making himself immediately available to help me. Meeting Nuno to present my academic work and gather his feedback, now as an expert in this area, brought me back precious memories from when we were going together through the hard work of all the group projects back in our telecommunications engineering course at ISEL. Finally, my old colleague João Tavares who worked with me in a company called Sol-S, 20 years ago, who immediately put me in touch with one of the most experienced cybersecurity consultants he has working in his company, who was also fantastic and extremely helpful to me in the discussion phase.

At last, but surely not least, I'd also like to thank my thesis advisor, professor Vitor Santos for his constant support, guidance, and availability to help me drive this thesis towards a successful conclusion. Professor Vitor always shown me the right way to take my work forward, helped me whenever I was feeling unsure and gave me confidence to feel I was going in the right direction. I'm happy for having decided to ask Professor Vitor to be my advisor and very grateful that he accepted it.

Abstract

One of the biggest concerns with the internet activity nowadays has to do with the increase of cybercrime. Cybercrime is one of the biggest threats for every organization in the world and its importance for society is reflected in numbers that show the impact it has in the world economy. Cybersecurity ventures predicted that by 2021 cybercrime will cost the world $6 trillion annually. This represents the greatest transfer of economic wealth in history, and it will become more profitable than the global trade of all major illegal drugs combined. The cyberattacks are getting more evolved every day and they leverage newer technologies, like artificial Intelligence (AI), to remain ahead of the curve and defeat the security measures that organizations put in place. To remain secure, organizations need to evolve faster and develop also defensive measures based on AI, otherwise they cannot remain effective, and ensure they are compliant with the security controls defined in the existing security frameworks. Studies demonstrate that there are some gaps in the existing security architectures that use artificial intelligence, and with this work one has developed a study in this area and built a security model, using the design science research methodology, leveraging AI technologies, to fill some of the identified gaps. This study aims to contribute to the industry, academia and to professionals working in the cybersecurity area, by providing a reference model that can be used to help choosing a defense strategy using AI techniques. The framework that was built in this work has been evaluated by cybersecurity experts which validated its utility, technical correctness and recognized that it brings a positive contribution in this space.

KEYWORDS

Cybersecurity, Cyberattacks, Malware, Information Security, Artificial Intelligence, Machine Learning Deep Learning

INDEX

FIGURE INDEX

TABLE INDEX

LIST OF ABBREVIATIONS AND ACRONYMS

A&R	Automation & Robotics
ACL	Access Control List
AI	Artificial intelligence
AIS	Artificial Immune System
ANN	Artificial Neural Network
ATD	Advanced Threat Response
BART	Bayesian Additive Regression Trees
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CART	Classification and Regression Trees
CASB	Cloud Access and Security Brokers
CC	Common Criteria
CIA	Confidentiality, Integrity, and availability
CNCS	Centro Nacional de Cibersegurança
CPS	Cyber Physical Systems
CVSS	Common Vulnerability Scoring System
DBFW	Database Firewall
DBN	Deep Belief Network
DDoS	Distributed Denial of Service
DL	Deep learning
DLP	Data Loss Prevention
DNS	Domain Name System
DoD	Department of Defence
DoS	Denial of Service
DRL	Deep Reinforcement Learning
DSRM	Design Science Research Methodology
DT	Decision Tree

EDR	Endpoint Detection and Response
FIC	File Integrity Checker
FIPS	Federal Information Processing Standards
FLS	Fuzzy Logic Systems
FQDN	Fully Qualified Domain Name
GA	Genetic Algorithm
HIC	Host-Based Intrusion Control
HIPS	Host-Based Intrusion Prevention System
IA	Information Assurance
IAM	Identity and Access Management
ICMP	Internet Control Message Protocol
ICT	Information and communications technologies
IDS	Intrusion Detection System
IEC	International Electrotechnical commission
InfoSec	Information Security
IOT	Internet of Things
IP	Internet Protocol
IPS	Intrusion Protection System
ISMS	Information Security Management System
ISO	International Organization for Standardization
IT	Information Technology
ITSEC	Information Technology Security Evaluation
KPI	Key Performance Indicator
LR	Linear Regression
ML	Machine Learning
MTD	Mobile Threat Detection
MV	Machine Vision

NB	Naïve Baines
NFW	Network Firewall
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NPB	Network Packet Broker
OSI	Open Systems Interconnection
PKI	Public Key Infrastructure
POD	Ping of Death
PRISMA	Preferred Reporting Items for Systematic Reviews and META-Analysis
RF	Random Forest
RFC	Request for Comments
RL	Reinforcement Learning
SEG	Secure Email Gateway
SIEM	Security Information and Event Management
SOC	Security Operations Centre
SOAR	Security Orchestration Automation and Response
SQL	Structured Query Language
SVM	Support Vector Machine
SWG	Secure Web Gateway
TCP	Transmission Control Protocol
TCSEC	Trusted Computer Security Evaluation Criteria
UDP	User Datagram Protocol
URL	Uniform Resource Locator
US	United States
VLAN	Virtual Local Area Network
XSS	Cross Site Scripting

1. INTRODUCTION

1.1 BACKGROUND AND PROBLEM IDENTIFICATION

Nowadays Cyber Security is a critical subject for individuals, enterprises, governments, and any type of organization. We're living in a world where everything is on the internet, from personal information and devices to enterprise systems and data, and even nations infrastructures, ensuring that data and infrastructure remains safe is one of the biggest challenges of Cybersecurity (Upadhyay, 2020).

Since nowadays all critical infrastructures are operating in the cyberspace, inevitably all are vulnerable to cyberattacks, meaning that, to keep these infrastructures and all their components operational and secure, information security becomes the most crucial aspect of the overall security of these types of systems. This means that the security of critical infrastructures is directly related to the information security measures that are implemented, and their effectiveness (Mikhalevich & Ryjov, 2018).

In the past few years, we have seen in the cyberspace a considerable growth of a variety of "smart" things and several other type of components and systems that can be targets for cyberattacks. The potential for these attacks to create a negative impact on information systems, automated control systems, data communications and telecommunications networks and other elements of critical infrastructures have grown significantly (Mikhalevich & Ryjov, 2018).

Moreover, the Cyberspace is also now considered a new war domain, it can even be considered the fifth dimension of warfare, to the point that all over the world, national security agencies are empowered to address cybersecurity. This is due to the evolution and increase of cybercrime and cyberthreats, which became the biggest threat to every company in the world having a huge impact on nations' economies. Over the past decade there have been several examples of cyberattacks on nations physical infrastructures, some of them resulted in military retaliation (Moura & Trilho, 2021).

Due to all this, nowadays all IT systems and infrastructures are required to comply with a set of security requirements to ensure the best protection against cyberattacks and other security threats. These requirements are based in the security controls defined by standards and best practices such as the ISO 27000 family of standards (ISO and IEC 2018), NIST guidelines and others. Focusing on ISO 27000 family we identify a wide range of technical and administrative controls for the avoidance, detection and mitigation of security risks (ISO/IEC, 2013).

Organizations, governments, and other public entities need to be concerned with the protection of all layers of their infrastructure that can serve as exploit vectors for these types of attacks. Vulnerabilities that can be exploited are present at all levels of an information system infrastructure, starting from the hardware layer, then the software defects, network infrastructure and protocols vulnerabilities, data and application layer (Jang-Jaccard & Nepal, 2014). The users in an organization are also a very critical aspect when it comes to consider them as a possible attack vector, since they are subject to fall into spam or phishing scams very easily. The wide usage of social media by the population makes them very exposed to cyberattacks. People share huge amounts of data on social media, sometimes personal data that can be used by criminals to perform phishing attacks, using social engineering techniques, that can easily lead to a more distracted user to install a piece of malware, either on his work computer, work mobile phone, or similarly on their personal devices (Jang-Jaccard & Nepal, 2014). Surely when doing so in a work computer or mobile phone, the

organization becomes vulnerable and the impact of such can be devastating. So, when adopting a strategy for cyberattacks protection, this variable must come into play.
It is hard for the existing security technologies in the market to keep up in a way to remain fully effective to detect and prevent these threats as much as possible (Tahir, 2018).
Attackers are evolving their methods very fast, and constantly improving their attacks strategy as well as the tools used. The old strategies of defense, where organizations used to invest in several types of inter-connected security solutions to protect their on-premises infrastructures, or even cloud platforms are no longer enough to prevent the new types of cyberattacks (Scott, 2017). This strategy of using several types of equipment's only introduce a bigger attack surface, leading to more points of potential vulnerabilities. Even if we are talking about modern solutions that can remain continuously and automatically updated, they cannot fully protect against the more evolved types of attacks (Scott, 2017).
These most evolved cyberattacks involve the utilization of Artificial Intelligence (AI). The usage of AI-driven techniques in the attack process, which are also called AI-based cyberattacks are sometimes done in conjunction with conventional attack techniques, resulting in more damage inflicted (Kaloudi & Jingyue, 2020).
According to (Kaloudi & Jingyue, 2020) there have been several recent studies on AI and cybersecurity, but researchers have not performed enough studies on AI-Driven cyberattacks in order for this phenomenon to be fully understood to the extent of being able to develop proper defenses against such attacks. (Kaloudi & Jingyue, 2020) have done some investigation of existing studies on this subject and mapped them into a framework providing insights to new threats. This framework provides a classification of malicious usage of AI during cyberattacks and provides basis for their detection with the objective to help predict future threats. However, during the study some limitations have been identified, like the lack of mitigation approaches.
(Chan et al., 2019) states that most research so far have focused on supervised learning of AI to teach an AI device how to protect systems mainly from malware and other external threats. However, this type of AI learning has trouble when it comes to handle new exploits and attacks.
From this, one concludes that there is still a lot of ground to be covered in this space when it comes to understand the evolution of cyber defence in terms of dealing with AI driven threats.

1.2 Study Objectives

The research goal is to build an effective conceptual security model with various AI based methods to help organizations improve their security architectures and be fully compliant with existing security standards like ISO 27002 or NIST cybersecurity framework. A model which can be used to improve the methods used to protect, detect, and respond to threats and attacks using AI driven technology.
To achieve this goal, the following objectives were defined:

- What is the current status of research in this area?
- What are the major issues of AI in cybersecurity for businesses and governments?
- What kind of AI techniques are currently useful in this area?
- What are the advantages and disadvantages of applying AI techniques in this field?

- Build a security intelligence model with various AI based methods.
- Validate the model.

1.3 Study Relevance and Importance

As addressed in the previous background analysis, the studies that have been done so far seem to have identified the lack of proper mitigation approaches when it comes to defend critical infrastructures from AI driven cyberattacks. Since this situation is apparently taking up some speed, it reveals that in order to prepare companies, public organizations and even governmental entities to be prepared for this growing scenario in the cyberspace, the security professionals need to become better prepared with tools that will help them mitigate threats and attacks of this type.

As an evident example of the relevance of this topic we can refer to (Şeker, 2019) where it is identified how the United States department of defense has framed this topic in the overall thematic of cybersecurity. According to (Şeker, 2019) amongst the five main pilars issued by the US department of defense on their cyberstrategy document, we can find the following: "*Advantages of technological change have to be preserved, developed, and new technologies (Especially Artificial Intelligence) have been adapted for cyber defense*" (Şeker, 2019). It goes on by stressing that this pilar is of particular importance, since "*Artificial intelligence became an indispensable element of cyber defense*" (Şeker, 2019). Understanding and implementing strategies and methods for cyber defense using artificial intelligence is now something attracting serious interest from researchers who are focusing on developing AI solutions and applications to protect critical infrastructure (Şeker, 2019).

The study performed by (Sarker et al., 2021) revealed a set of gaps in this area, which can be used as a research direction. As an example it has been identified that a gap that can potentially result in a research area is the design of security ontologies according to today's requirements and known representation models in order to build effective conceptual security framework (Sarker et al., 2021). Moreover (Sarker et al., 2021) identifies the need to "*build an effective cyber security framework that supports artificial intelligence*" , "*the most important task for an intelligent cyber security system*".

With this study one expects to contribute to the industry, academia and to professionals working in the cybersecurity area, with a reference model which can be used to help them chose a defense strategy using AI techniques and fill the identified gap in this space.

1.4 Document Structure

This thesis document is structured in the following way:

The first chapter, as presented so far, addresses a brief introduction of the studied topic providing its background and identifying the problem that one has proposed to study and research. Additionally, it presents the objectives of this study and describes the importance and relevance of conducting such a study. Chapter two is focused on the methodology used in this research. One used the DSRM methodology, and in this chapter, one describes the methodology and how it was used to conduct this work. In chapter three one performs a literature review on the two main topics in research -

Cybersecurity and Artificial intelligence. Each of the topic is studied in detailed, focusing on the following aspects:

Cybersecurity – Analyze and understand the main concepts, the relevance and impact of cybersecurity in today's economy, what are the existing security risks, threats and types of attacks most used, understand existing security standards in the industry, and with this, identify challenges and opportunities that arise from this analysis.

Artificial Intelligence – Also very important to start with an analysis of the main concepts of this type of technology, then identify the areas where it's used, focus on the specific aspect of machine learning, which plays a critical role in the artificial intelligence landscape, and finally understand how artificial intelligence and cybersecurity link together, with the purpose to identify how AI is today used in cybersecurity, identifying it's benefits and also its drawbacks and limitations, so one can build the model from there.
One finishes this chapter by describing the methodology used for the systematic literature review: the PRISMA methodology, to provide the answers to the research questions:

- What is the current status of research in this area?
- What are the major issues of AI in cybersecurity for businesses and governments?
- What kind of AI techniques are currently useful in this area?
- What are the advantages and disadvantages of applying AI techniques in this field?

Chapter four is about building the framework for one's model, perform its validation, discussion of the output and results, leading to a revised and final model implementation. In chapter five one presents his conclusions on this work, providing a brief synthesis of the work done, identifying the research limitations and leave some proposals for future work in this area, based on found limitations.

2 METHODOLOGY

The objective of this study, as mentioned above, is to build a conceptual security model based on AI technologies. In this sense the proper research methodology to use is the Design Science Research methodology (DSR). Design science is about the way we can combine several components to create an artifact. A conceptual model is an artifact, and this artifact is built with the objective of solving a particular problem or improve upon existing solutions (Simon, 1996). By building this model we are creating something that can be instantiated and then evaluated, which, by definition, falls in the design science research methodology.

2.1 DESIGN SCIENCE RESEARCH (DSR)

The definition of DSR as per (Hevner & Chatterjee, 2010) is "*a research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing with new knowledge to the body of science evidence. The design artifacts are both useful and fundamental in understanding that problem.*"

One has found several literature about DSR methodology, all of them proposing similar models to develop artifacts that solve a particular problem, but for the purpose of this work this researcher will follow an adaptation of the model proposed by (Peffers et al., 2007). This model consists of several iteration steps which are used while applying the methodology, starting by identifying the problem and its motivation, define the objectives of the solution, then design and develop the solution, evaluate it, and finally communicate it as a new disciplinary knowledge for the scientific and academic community. The model adaptation and its steps can be depicted like in figure 1.

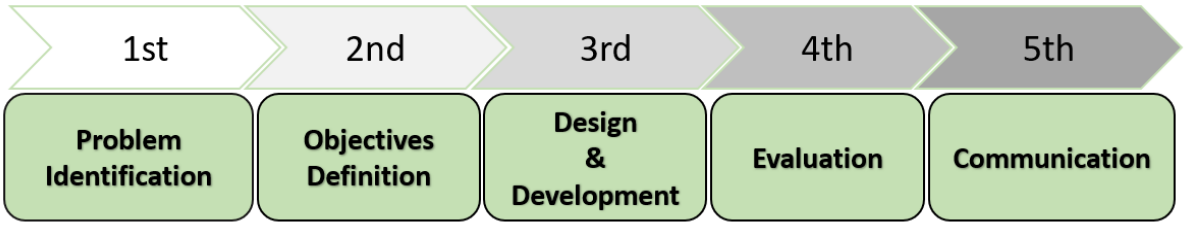

Figure 1 - DSRM Process Model Adaptation (Peffers et al., 2007)

Each of these iteration steps were adapted to achieve the goal of this research and can be briefly described as follow.

The first step is where the researcher defines his specific research problem and presents a justification for the value his solution will have. By justifying its value, the researcher achieves two critical goals, one is ensuring his own motivation to perform the work, and the second is the motivation of his audience to accept his results. The focus is on gathering knowledge on the state of the identified problem and the importance the proposed solution will have (Peffers et al., 2007).

From the identified problem and after gathering knowledge of what's possible, in the second step, the researcher can then define the objectives for his proposed solution. These objectives are set in terms of what it will bring in terms of additional benefits compared to existing solutions, or how it will support other solutions for the same problem. This can only be achieved after having acquired

considerable knowledge on the state of the problem in study, and the existing solutions for it and their efficiency or lack of it (Peffers et al., 2007).
The next step, the 3rd iteration, is about the development of the solution itself, i.e., the creation of the artifact. This artifact can be a model, a method, or an instantiation of it. In this iteration the researcher determines the artifacts architecture it's desired functionality which then leads to the creation of the artifact itself (Peffers et al., 2007).
In the 4th step, the researcher shall evaluate the artifact, by observing and measuring how well the solution supports the resolution of a specific problem. At this point one shall compare the objectives proposed for the solution with the results that come out of its utilization in a practical situation of solving one or more problems. Collecting feedback from subject matters experts or other relevant audience is also a valuable way of evaluating the applicability of the artifact. Depending on the evaluation result, one may have to go back to the previous step to make some corrections and adjustments which will result in a new redesigned artifact which surely will be more effective than the previous one (Peffers et al., 2007).
The 5th and final step of this process is the communication of the newly created artifact. It's importance, utility, effectiveness shall be communicated to other researchers, industry subject matter experts, academia, scholar publications and other parties of interest, who are considered to benefit from the new found knowledge (Peffers et al., 2007).

2.2 RESEARCH STRATEGY

The research strategy adopted aimed to be a practical implementation of the steps described previously. So, it all started with the problem identification, which can only be identified once the study topic is defined. Here was a matter of personal interest: cybersecurity, as the chosen discipline, triggered this researcher to perform brief research of the main issues that exists today which are related to the cybersecurity domain. Having realized that the utilization of AI is a very important aspect for today's cybersecurity and having learned that there is currently lots of research interest, as well as some identified problems related to the utilization of AI in the cybersecurity domain, the study topic has been defined. Still on the preliminary research on this specific topic, some literature was analyzed, and the specific problem chosen as the main target of this work has been identified. The problem identification and the motivation behind it, as well as its value and importance, has been addressed in section one of this work.
This first step allowed this researcher to gather preliminary knowledge on the existing solutions for the identified problem, highlighting the gap that one intends to fill with this work which is a more complete cybersecurity framework that uses several AI techniques. This then acted as a Segway to perform a deeper study of the two main topics, i.e., cybersecurity and AI, where both disciplines main concepts were studied in detail, followed by a systematic literature review about the current state of the art of the utilization of AI in the cybersecurity domain. In the systematic literature review is where the researcher crosses the information specifically related to the utilization of AI in cybersecurity. By performing this systematic literature review, which is thoroughly described in the next section of this work, this researcher was able to find the answers for the intermediate objectives and acquire a very good understanding of the state of the art of how AI is currently applied to cybersecurity, what are its benefits, which are the main challenges, where there is space to improve, where are the main gaps. Having done this, the researcher was able to set the objectives that would result in a proposed framework that covers the identified gaps, which is ultimately the goal of this

work. At the point of conclusion of the literature review, the researcher had acquired all the knowledge and information needed to set the specific objectives for the model to build, and as such, at this point the researcher moved to the design and development phase. The design and development of the model was a natural and smooth process because the foundations provided by the literature review conducted previously were solid. Having built the model, the researcher performed an evaluation by presenting the proposed model to subject matters experts, conducting several interviews and colleting their feedback. At this stage of the DSR methodology, it was not possible to perform an evaluation of the framework by measuring how well it supports the resolution of a specific problem so the chosen option at this 4th stage was to rely solely on the expert interviews. These expert interviews resulted in several suggestions for corrections and improvements, and based on these, some adaptations were performed, which resulted in a revised framework as the final proposal.

Unfortunately, due to time limitation, it wasn't possible to perform the 5th stage of the research strategy which is the communication of the work to the communities of interest. As such, this has been highlighted as a limitation of the work.

3 LITERATURE REVIEW

A good literature review is not just a collection of summaries of scientific papers or other articles together with a bibliography list (Levy, Yair;Ellis, 2012).
The objective of performing a systematic literature review is to gather information from several scientific sources on the subjects in study, synthetize it and gather a new perspective, discover the most important variables, identify relationships between ideas and practices, have a general overview and establish a context of the several topics that are relevant within the wider main subjects of the study: in this case Cybersecurity and AI, with the purpose of justifying a particular approach, corroborate theories or, oppositely disprove them (Hart, 1988).
(Liberati et al., 2009) describes systematic reviews as an essential tool to help researchers summarize evidence accurately and reliably. It "*attempts to collate all empirical evidence that fits pre-specified eligibility criteria to answer a specific research question*" (Liberati et al., 2009).
It's essential that in every academic project one conducts an effective review, which allows the researcher to create a solid foundation to contribute to knowledge development. By doing so, theory development is facilitated, it identifies areas where the research already exists and uncovers other areas where new research is required (Webster & Watson, 2002).
Based on these principles, one has conducted structured research using the available sources, to collect and explore relevant literature that could support this study. With it one was able to consolidate knowledge on the subject, identify research gaps, find the answers for the research questions, and with all this, build the knowledge foundation to develop the architecture framework/model proposed, which is presented in chapter four.
In the next sections, the acquired knowledge which resulted from this review, is presented.

3.1 CYBERSECURITY

One started this study by asking the question: what is cybersecurity? Having analyzed several literature on this topic one could conclude that this term is broadly used, there are several different definitions which vary according to context, these are found to be often subjective, sometimes also uninformative (Craigen et al., 2014).
This broad and subjective definition lead (Craigen et al., 2014) to conduct a study where they have worked towards a better definition of cybersecurity which resulted in "*Cybersecurity is the organization and collection of resources, processes, and structures used to protect cyberspace and cyberspace-enabled systems from occurrences that misalign de jure from de facto property rights*".
While conducting this study (Craigen et al., 2014) chose to list some other definitions found in literature, which one finds relevant to describe here in the context of this work. These are:

"*Cybersecurity consists largely of defensive methods used to detect and thwart would-be intruders*." (Kemmerer, 2003)

"*Cybersecurity entails the safeguarding of computer networks and the information they contain from penetration and from malicious damage or disruption*." (Lewis, 2006)

"*Cyber Security involves reducing the risk of malicious attack to software, computers and networks. This includes tools used to detect break-ins, stop viruses, block malicious access, enforce authentication, enable encrypted communications, and on and on*." (Amoroso, 2006)

"*Cybersecurity is the collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance and technologies that can be used to protect the cyber environment and organization and user's assets*."(ITU, 2008)

"*The ability to protect or defend the use of cyber- space from cyber-attacks*." (CNSS, 2010)

"*The body of technologies, processes, practices and response and mitigation measures designed to protect networks, computers, programs and data from attack, damage or unauthorized access so as to ensure confidentiality, integrity and availability*." (Ministry of Public Safety, 2010)

"*The art of ensuring the existence and continuity of the information society of a nation, guaranteeing and protecting, in Cyberspace, its information, as- sets and critical infrastructure.*" (Canongia, C ; Mandarino, 2014)

"*The state of being protected against the criminal or unauthorized use of electronic data, or the measures taken to achieve this.*" (Oxford University Press, 2014)

"*The activity or process, ability or capability, or state whereby information and communications systems and the information contained therein are protected from and/or defended against damage, unauthorized use or modification, or exploitation.*" (DHS, 2014)

Another study conducted by (Peslak, Alan; Hunsinger, 2019), lead to the following definition "*Cybersecurity, in general, is the protection of data, information, devices, and systems from unauthorized or malicious attacks or access*".

One more important definition of cybersecurity is provided by Kaspersky Labs, an industry renowned software vendor in this area "*Cyber-security is the practice of defending computers, servers, mobile devices, electronic systems, networks, and data from malicious attacks. It's also known as information technology security or electronic information security. The term applies in a variety of contexts, from business to mobile computing, and can be divided into a few common categories*" (Kaspersky Labs, 2019).

3.1.1 Concepts

Following an initial approach towards the several definitions of cybersecurity, we need to understand what these definitions mean. In a more simplistic way, we can say that cybersecurity is about securing things that are vulnerable through information and communication technologies (ICT).
This has the main objective of defending data from malicious attacks, i.e., ensure the security of stored data and information, and also the technologies used to secure that data and information (Amit, 2016). On this note is relevant to highlight that Cybersecurity and Information security are

seen in distinct ways, although they often are used as synonyms (Amit, 2016). One will go into these differences later in this text.
Data security is about securing data, and there is also a difference between data and information. Not all data is information, data can be called information if it can be interpreted and if it has a meaning within a certain context. As an example, "01111755" is data, but if we know or if we interpret it as the date 1 of November 1755, we will recognize it as a date when a big earthquake took place in Lisbon. This now becomes information. Information can therefore be defined as data that has some meaning (Amit, 2016).
One has said that information security and cybersecurity are seen in distinct ways, in fact they're two different things.
As seen previously, there are several definitions of cybersecurity, and the same applies for information security, but for the sake of this comparison we will rely on the following definitions as set by NIST.

Information Security – "*The protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability*" (NIST, 2013).

Cybersecurity – "*The ability to protect or defend the use of cyberspace from cyber-attacks*" (NIST, 2013).

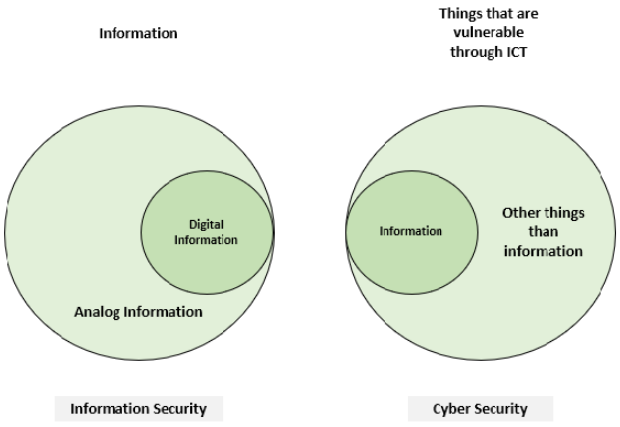

Figure 2 – Information Security vs Cybersecurity (Amit, 2016)

In using NIST definition of cybersecurity it has been introduced in this text a new term called cyberspace. Cyberspace is, also according to NIST definition, "*a global domain within the information environment consisting of the interdependent network of information systems infrastructures including the internet, telecommunications networks, computer systems, and embedded processors and controllers*" (NIST, 2013).
The information that we aim to secure when using cybersecurity has some critical characteristics, or properties, that are crucial to maintain intact, as information value comes from the properties it possesses. When a property of the information changes its value either increases or, more

commonly, decreases. Each of these characteristics, or properties, are represented by what it's called the CIA triad (NIST, 2013):

Figure 3 – The CIA Triad (NIST, 2013)

Confidentiality – The property that information is not made available or disclosed to unauthorized individuals, entities, or processes.

Integrity – The property that data has not been altered or destroyed in an unauthorized manner.

Availability – The fact that data is accessible, and services are operational, meaning ensuring timely and reliable access to and use of information.

Other important properties of information, although not part of the CIA triad are (NIST, 2013):

Accuracy – The property that Information is free from mistakes or errors, and it has the value that the end user expects.

Authenticity – The property that information is the quality or state of being genuine or original, rather than a reproduction or fabrication.

Utility – The quality or state of having value for some purpose or end.

Possession – The possession of information is the quality or state of ownership or control.

Based on the mentioned definitions, it can be said that cybersecurity is concerned with the security of anything in the cyber realm, whilst information security is concerned with the security of information regardless of the realm, figure 4. So, from this, it can also be said that information security is a super set of cybersecurity (NIST, 2013).

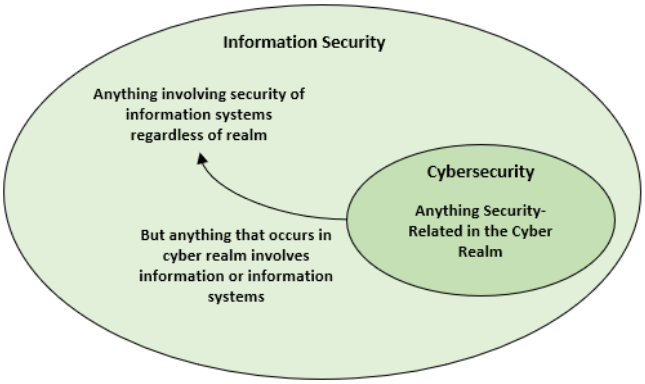

Figure 4 – Information Security and Cybersecurity Realms (NIST, 2013)

In figure 2 we can see that the image on the right represents the Cybersecurity, i.e., things that are vulnerable through ICT, which includes information, both physical and digital, and non-information such as facilities, cars, traffic cameras and other type of IOT devices. The image on the left represents the information security, i.e., information both digital and analog.

IT security, which is also a new term in this text, is the protection of information technologies. Having defined that, it can be safely said that when comparing ICT security and IT security there are no differences (Amit, 2016).

One can overlay the two images from figure 2 and generate a diagram depicting the relationship between ICT security, cyber security, and information security.

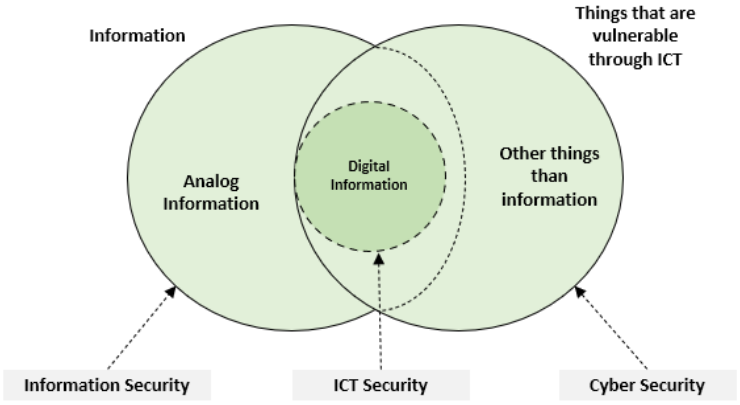

Figure 5 – Information Security, ICT Security and Cybersecurity (Amit, 2016)

As seen previously, cybersecurity includes everything that can be reached in the cyberspace. This might lead us to the idea that everything in the world is vulnerable through ICT, however this discipline and its definition teach us that we should focus on protecting only what's to be protected based in the security challenges presented by the utilization of ICT (Amit, 2016).

So, what needs to be protected then? With the huge proliferation of data, several types of control devices and the growth of Internet of Things (IoT), the systems that need to be protected include all connected devices, as well as all its control systems and of course data. Any type of computers or servers, any type of mobile devices, networks devices, and databases all fall under this umbrella. Also, all electronic and non-electronic systems that control any device like cars, RFID chips, smart

home devices (electronic appliances), house alarms, and so on, need protection from malicious attacks. All this is part of the vast world of cybersecurity today (Peslak, Alan; Hunsinger, 2019). (Kaspersky Labs, 2019) lists some critical domain areas that are included in cybersecurity:

Network security - The activity of securing a computer network from unauthorized access (intruders), cyber attackers, malware, and other types of threats.

Application security – Is about making sure we keep software and devices free of threats. If an application gets compromised attackers will get access to data which is supposed to be protected by the application itself. When designing an application, we must ensure the security is part of the design phase, even before the application is built and deployed.

Information security – As described before, protects the confidentiality, integrity, and availability of data, both at rest and in transit.

Operational security – It's all about the necessary processes and activities that need to be performed for correctly handling and protecting data assets. Define the right permissions users need when accessing a network or any system, and the procedures that determine how and where data can be stored or shared are examples of what falls into this domain.

Disaster recovery and business continuity are two critical domains which define how an organization responds to a cybersecurity incident or any other event that causes the loss of operations or data:

Disaster recovery – Is comprised of a set of policies that define how an organization can restore its systems operations and information, to return to the same operating capacity as before the security incident, or any other disruptive event.

Business continuity - Is a fall-back plan defined by the organization, which allows it to continue its operations even if certain resources are unavailable due to a security incident or any other similar disruptive event.

End-user education – Is about dealing with people, which sometimes can be the most unpredictable cyber-security factor. A user unaware of the right security behaviours and practises can accidentally introduce a virus or other type of malware into a system that was considered secure until then. Security awareness training for users, in topics like phishing emails, educating them to delete suspicious email attachments for example, or not plug in unidentified USB drives into their work computers, and several other relevant learnings is crucial for keeping any organization secure.

To finalize this section about cybersecurity concepts, one describes a set of terminology used in ICT security. These concepts are crucial to understand, as later in the text one will refer to them in the context of its use and or impact on the overall cybersecurity landscape. This terminology can be found in Network working group request for comments 4949 (Shirey, 2007):

Adversary (Threat agent)

- "*An entity that attacks or is a threat to a system*".

Attack

- "*An assault on system security that derives from an intelligent threat; that is, an intelligent act that is a deliberate attempt (especially in the sense of a method or technique) to evade security services and violate the security policy of a system*".

Countermeasure

- "*An action, device, procedure, or technique that reduces a threat, a vulnerability, or an attack by eliminating or preventing it, by minimizing the harm it can cause, or by discovering and reporting it so that corrective action can be taken*".

Risk

- "*An expectation of loss expressed as the probability that a particular threat will exploit a particular vulnerability with a particular harmful result*".

Security Policy

- "*A set of rules and practises that specify or regulate how a system or organization provides security services to protect sensitive and critical system resources*".

System Resource (Asset)

- "*Data contained in an information system; or a service provided by a system; or a system capability, such as processing power or communication bandwidth; or an item of system equipment (i.e., a system component – hardware, firmware, software, or documentation); or a facility that houses system operations and equipment*".

Threat

- "*A potential for violation of security, which exists when there is a circumstance, capability, action, or event, that could breach security and cause harm. That is, a threat is a possible danger that might exploit a vulnerability*".

Vulnerability

- "*A flaw or weakness in a system's design, implementation, or operation and management that could be exploited to violate the system's security policy*".

3.1.2 Security Functional requirements

Security functional requirements are a list of controls and assurance recommendations covering seventeen security-related areas with the objective of assuring the protection of confidentiality, integrity and availability of data and information systems. More specifically in this text one is referring to the FIPS 200 recommendations, which has been defined by NIST to protect the United Sates federal information and information systems (NIST, 2006).

The controls and recommendations on this list aim to be in fact countermeasures used with the objective of reducing vulnerabilities and handle threats to system assets.

Organizations and individuals value their systems and data (assets), so they wish to minimize risk to it by imposing countermeasures to protect these assets from threats that are raised by threat agents who wish to inflict damage.

This workflow can be represented in a picture like the one below:

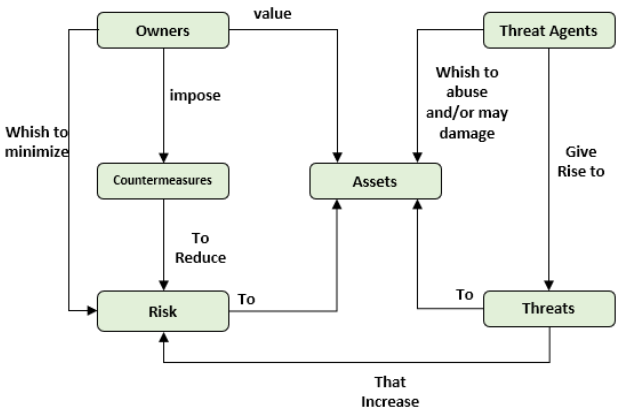

Figure 6 – Asset Security workflow

According to this publication, organizations must meet these minimum-security functional requirements to be compliant. The seventeen security related areas and its recommendations are (NIST, 2006):

1. **Access control (AC)**

"*Organizations must limit information system access to authorized users, processes acting on behalf of authorized users, or devices (including other information systems) and to the types of transactions and functions that authorized users are permitted to exercise*".

2. **Awareness and training (AT)**

"*Organizations must: (i) ensure that managers and users of organizational information systems are made aware of the security risks associated with their activities and of the applicable laws, Executive Orders, directives, policies, standards, instructions, regulations, or procedures related to the security of organizational information systems; and (ii) ensure that organizational personnel are adequately trained to carry out their assigned information security-related duties and responsibilities*".

3. **Audit and accountability (AU)**

"*Organizations must: (i) create, protect, and retain information system audit records to the extent needed to enable the monitoring, analysis, investigation, and reporting of unlawful, unauthorized, or inappropriate information system activity; and (ii) ensure that the actions of individual information system users can be uniquely traced to those users so they can be held accountable for their actions*".

4. Certification, accreditation, and security assessments (CA)

"*Organizations must: (i) periodically assess the security controls in organizational information systems to determine if the controls are effective in their application; (ii) develop and implement plans of action designed to correct deficiencies and reduce or eliminate vulnerabilities in organizational information systems; (iii) authorize the operation of organizational information systems and any associated information system connections; and (iv) monitor information system security controls on an ongoing basis to ensure the continued effectiveness of the controls*".

5. Configuration management (CM)

"*Organizations must: (i) establish and maintain baseline configurations and inventories of organizational information systems (including hardware, software, firmware, and documentation) throughout the respective system development life cycles; and (ii) establish and enforce security configuration settings for information technology products employed in organizational information systems*".

6. Contingency planning (CP)

"*Organizations must establish, maintain, and effectively implement plans for emergency response, backup operations, and post-disaster recovery for organizational information systems to ensure the availability of critical information resources and continuity of operations in emergency situations*".

7. Identification and authentication (IA)

"*Organizations must identify information system users, processes acting on behalf of users, or devices and authenticate (or verify) the identities of those users, processes, or devices, as a prerequisite to allowing access to organizational information systems*".

8. Incident response (IR)

"*Organizations must: (i) establish an operational incident handling capability for organizational information systems that includes adequate preparation, detection, analysis, containment, recovery, and user response activities; and (ii) track, document, and report incidents to appropriate organizational officials and/or authorities*".

9. Maintenance (MA)

"*Organizations must: (i) perform periodic and timely maintenance on organizational information systems; and (ii) provide effective controls on the tools, techniques, mechanisms, and personnel used to conduct information system maintenance*".

10. Media protection (MP)

"*Organizations must: (i) protect information system media, both paper and digital; (ii) limit access to information, on information system media to authorized users; and (iii) sanitize or destroy information system media before disposal or release for reuse*".

11. Physical and environmental protection (PE)

"*Organizations must: (i) limit physical access to information systems, equipment, and the respective operating environments to authorized individuals; (ii) protect the physical plant and support infrastructure for information systems; (iii) provide supporting utilities for information systems; (iv) protect information systems against environmental hazards; and (v) provide appropriate environmental controls in facilities containing information systems*".

12. Planning (PL)

"*Organizations must develop, document, periodically update, and implement security plans for organizational information systems that describe the security controls in place or planned for the information systems and the rules of behaviour for individuals accessing the information systems*".

13. Personnel security (PS)

"*Organizations must: (i) ensure that individuals occupying positions of responsibility within organizations (including third-party service providers) are trustworthy and meet established security criteria for those positions; (ii) ensure that organizational information and information systems are protected during and after personnel actions such as terminations and transfers; and (iii) employ formal sanctions for personnel failing to comply with organizational security policies and procedures*".

14. Risk assessment (RA)

"*Organizations must periodically assess the risk to organizational operations (including mission, functions, image, or reputation), organizational assets, and individuals, resulting from the operation of organizational information systems and the associated processing, storage, or transmission of organizational information*".

15. Systems and services acquisition (SA)

"*Organizations must: (i) allocate sufficient resources to adequately protect organizational information systems; (ii) employ system development life cycle processes that incorporate information security considerations; (iii) employ software usage and installation restrictions; and (iv) ensure that third-party providers employ adequate security measures to protect information, applications, and/or services outsourced from the organization*".

16. System and communications protection (SC)

"*Organizations must: (i) monitor, control, and protect organizational communications (i.e., information transmitted or received by organizational information systems) at the external boundaries and key internal boundaries of the information systems; and (ii) employ architectural designs, software development techniques, and systems engineering principles that promote effective information security within organizational information systems*".

17. System and information integrity (SI)

"*Organizations must: (i) identify, report, and correct information and information system flaws in a timely manner; (ii) provide protection from malicious code at appropriate locations within organizational information systems; and (iii) monitor information system security alerts and advisories and take appropriate actions in response*".

3.1.3 Security Standards & Frameworks

Before going into the details of the existing security standards and frameworks in the industry, explain what they are and what they contain, one will start by describing a few other important concepts as a baseline for this subject. Cybersecurity, as previously addressed, is about information assurance, so we need to understand this concept also, and as a basis we shall define the concept of security, risk, and risk management. According to (Bosworth et al., 2009) Security can be defined as "*the state of being free from danger and not exposed to damage from accidents or attack, or it can be defined as the process for achieving that desirable state*". When focusing on information systems security, the goal is to optimize the performance of an organization with regards to the risks to which it is exposed. Risk can also be defined as "*the chance of injury, damage or loss*", and it has two elements: chance which is an element of uncertainty, and potential loss or damage (Bosworth et al., 2009).

Risk management can be defined as the "*optimization process of resource allocation, by minimizing the total cost of information system security measures taken, and the risk losses experienced"* (Bosworth et al., 2009). Risk management can be seen as a three part process: i) Identification of material risks; ii) Selection and implementations of measures to mitigate the risks; iii) Tracking and evaluating risk losses experienced, in order to validate the first two part of the process (Bosworth et al., 2009). Knowing this we are now in position to provide a definition of information assurance. It can be defined as "*the practice of assuring information and managing risks related to the use, processing, storage, and transmission of information*" (Sosin, 2018). In fact, Information assurance includes protection of the integrity, availability, authenticity, non-repudiation and confidentiality of user data (Sosin, 2018).

Security standards and frameworks in essence are models created by the industry to help organizations implement the best practises, rules, and guidelines to ensure information assurance.

More specifically, a cybersecurity framework contains processes, practises, and technologies. A cybersecurity standard contains statements that describe what must be achieved in terms of security outcomes to fulfil an enterprise's stated security objectives (Bosworth et al., 2009).
There are several types of standards which are pertinent to security. A few examples, not exhaustive, are (Bosworth et al., 2009):

- Capability standards
- Personnel certifications
- Risk assessment criteria
- Requirement's specifications
- Functional specifications
- Assurance specifications
- Performance criteria
- Product development standards
- Testing, evaluation, and assessment standards/criteria
- Product review criteria
- Interoperability standards
- Procurement standards
- Ancillary standards

Regarding frameworks, there are also some examples which are pertinent to security (Bosworth et al., 2009):

Framework	Brief description
ISO 27001	A set of standards to put managers in control of the cybersecurity measures.
GDPR	One of the latest frameworks legislated to secure personally identifiable information belonging to European citizens.
COBIT	An ISACA framework that integrates a business's best aspects to its IT security, governance, and management.
PCI DSS	Applies to companies that handle credit card information.
CIS Critical Security Controls	A set of 20 actions designed to mitigate the threat of most common cyber-attacks.
NIST Cybersecurity	General-use framework to support organization's cybersecurity.
NIST SP 800-53	A catalogue of security and privacy controls for all U.S. federal information systems.
Zero Trust	Is a security concept centred on the belief that organizations should not automatically trust anything inside or outside.
Gartner DSG	A private Data Security Governance framework from Gartner.

Table 1 – Cybersecurity Frameworks

On the standards is of utmost importance to reference the security standards for products.
The Trusted Computer Security Evaluation Criteria (TCSEC), also known as orange book, was released in 1985 and developed by the US department of defence (DoD). This standard was developed mainly focused on military applications (DoD, 1985). This standard has been used until 2000 and since then replaced with a new one called Common Criteria (CC). The Information Technology Security Evaluation (ITSEC), is the European analogous of the US's TCSEC, was developed by France, Germany, United Kingdom, and the Netherlands and was published in 1991. This standard was widely used also in the military sector and additionally for digital signatures. The ITSEC has also been replaced by the Common Criteria standard (CC). The CC is a common set of criteria for evaluating the security of computer systems, developed by the national security authorities of USA, Canada and Europe (CCRA, 1996):

- Product evaluation by independent, licensed laboratories
- Documents defining the certification process
- Certifications issued by Certificate Authorizing Schemes (subset of CCRA members)
- Certifications recognized by all CCRA members

The ISO 27001 Standard

ISO 27001 is a standard that belongs to the ISO/IEC 27000 family of standards, and it's composed of two parts: the mandatory controls and the optional controls. It is the international standard for information security, and it's meant to set out specifications for an information security management system (ISMS) (ISO/IEC, 2013). By adopting these best-practices approach, organisations will be better managing their information security by addressing people, processes, and technology. Organizations can obtain certification in ISO27001 and by doing so they can be recognised worldwide as an indication that their ISMS is aligned with information security best practices. ISO 27001 is a framework that helps organisations "*establish, implement, operate, monitor, review, maintain and continually improve an ISMS*"(ISO/IEC, 2013).

The NIST Cybersecurity Framework

This framework is one of the most important ones in the industry. NIST introduces this framework as a widely used approach to help determine and address highest priority risks to businesses, including standards, guidelines, and best practices. The NIST Framework is focused on "*using business drivers to guide cybersecurity activities and considering cybersecurity risks as part of any organization's risk management processes*" (Barrett, Matt, 2018). Although it's been developed to improve cybersecurity for US critical infrastructures, the framework "*can be used by organizations in any sector or community*" (Barrett, Matt, 2018). The framework is essentially composed of five core functions: Identify, Protect, Detect, Respond and Recover. The core is the highest level of abstraction of this framework, in each of the functions we have also several categories which in turn also have subcategories. Core functions helps organizations express their management of cybersecurity only at high level. In each of these core functions organizations should aim to ask the following questions (Barrett, Matt, 2018):

Identify: What processes and assets need protection?
Protect: What safeguards are available?
Detect: What techniques can identify incidents?
Respond: What techniques can contain impact of incidents?
Recover: What techniques can restore capabilities?

Figure 7 – NIST framework core functions (Barrett, Matt, 2018)

The answers for the questions posed in each of the functions result in several outcomes. As examples we have (Barrett, Matt, 2018):

Identify

- Identifying physical and software assets to establish an Asset Management program
- Identifying cybersecurity Policies to define a Governance program
- Identifying a Risk Management Strategy for the organization

Protect

- Establishing Data Security protection to protect the confidentiality, integrity, and availability
- Managing Protective Technology to ensure the security and resilience of systems and assists
- Empowering staff within the organization through Awareness and Training

Detect

- Implementing Security Continuous Monitoring capabilities to monitor cybersecurity events
- Ensuring Anomalies and Events are detected, and their potential impact is understood
- Verifying the effectiveness of protective measures

Respond

- Ensuring Response Planning processes are executed during and after an incident
- Managing Communications during and after an event
- Analysing effectiveness of response activities

Recover:

- Ensuring the organization implements Recovery Planning processes and procedures
- Implementing improvements based on lessons learned

- Coordinating communications during recovery activities

To move from a high-level view to a more granular exercise of cybersecurity management, organizations should also look at implementing actions as per the categories into each core function. The following picture provides a representation of some example categories that can be used (Barrett, Matt, 2018):

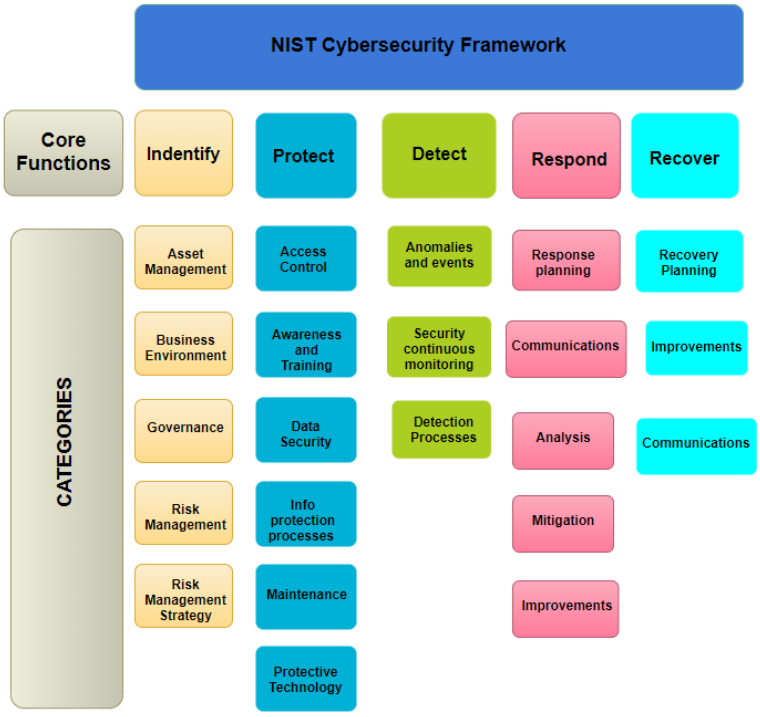

Figure 8 – NIST framework categories (Barrett, Matt, 2018)

3.1.4 Fundamental Security Design Principles

In the previous section one has said that security standards and frameworks are models which help organizations implement the best practises, rules, and guidelines to ensure information assurance. Models will therefore support the way organizations design their security. And to design their security there are a set of security design principles that should be followed. These principles are important to highlight some important concepts and to help infosec professionals to build an understanding of a complex reality. The first time that security principles were introduced was in 1975 when the document "*The Protection of Information in Computer Systems*" (Saltzer & Schroeder, 1975) was published, and this document defined a series of design principles to secure systems. As years have passed, systems and technology evolved, and in 2012 (Smith, 2012) wrote the article "*A Contemporary Look at Saltzer and Schroeder's 1975 Design Principles*" which compared those design principles with a new list which have been developed in the meantime, and also introduced a few more. (Saltzer & Schroeder, 1975) initially listed eight principles for computer security which are:

- **Economy of mechanism**: This principle is also known as “Keep it Simple Sir” (KISS), and the logic of it is that the more complicated something is, the greater the chance there is something wrong with it. “*A simpler design is easier to test and validate*”.

- **Fail-safe defaults**: In computer systems the safe default should generally be “no access”. The system must specifically grant access to resources. By default, we should deny everything and grant access to just the absolute necessary.

- **Complete mediation**: Access rights should be completely validated every time there is an access to a resource. Systems should rely as little as possible on access decisions retrieved from a cache. Trust by verify.

- **Open design**: Obscurity is not security. The rational of this is that the design itself is what provides security, not the secrecy of the design. Example is cryptography which uses open standards (AES, 3DES, etc). Everyone knows the standards but it’s how the cipher is designed that makes it secure.

- **Separation of privilege**: Also known as segregation of duties. Permissions to a system should be designed in a way that there is no single individual with access to everything. It’s using a concept of having more than one person required to perform a task. This prevents conflict of interests, wrongful acts, fraud, abuse, and errors. We can use as example the access to a nuclear missile, where it’s always required two keys to launch an attack.

- **Least privilege**: Based on the need-to-know logic. Any object relevant to a system, being a user, an administrator, a program, a system account or any other, should only have the privileges it needs to perform its task. As an example, a user should not login to his work computer with an administrator account. An application administrator should only have rights on the application he manages, and no other administrator rights on the system or other applications.

- **Least common mechanism**: Users should not share systems mechanisms unless necessary. Sharing resources provides a channel along which information can be transmitted.

- **Psychological acceptability**: Based on the logic that users will only adhere to a certain policy if they can understand its need. If we design security mechanisms that don’t make sense to users, they will find difficult to follow them, leading to potentially security issues.

On top of these 8 principles listed by (Saltzer & Schroeder, 1975), more recently (Smith, 2012) noted two additional principles that were well-known for physical security, which they’ve also recognized as relevant for computer security, and which are widely used today:

- **Work factor**: The stronger the security measures we implement, the harder the attacker will have to work to defeat them. Proper examples are long and more complex passwords, and larger encryption keys, meaning that an attacker will have to perform much more trial and error attempts until he can either discover a password or break a cypher.

- **Compromise recording**: Systems should keep a record of any attack, even if the attacks aren't blocked. This is an essential feature for logging and auditing. For each event that occurs there should be a record of who performed the action (user, system, or process), what was the action (description of the action taken), When did it occur (timestamp which should be synchronized across systems), where it happened (object involved or acted on to perform the action).

With time more principles emerged, and (Smith, 2012) introduced several other principles, like these ones:

- **Defence in depth**: One should build a system with independent security layers so that an attacker must defeat multiple security measures for the attack to succeed. It's the logic of a medieval castle, where the attackers needed to pass several layers of defence until they could enter the castle.

- **Transitive Trust:** If A trusts B and B trusts C, then A also trusts C. In a sense, this is an inverted statement of the least common mechanism. Transitive trust is currently a term widely used in computer security.

- **Continuous improvement**: Continuously assess how the security objectives are being achieved and make the necessary changes to improve the results. The standards that have been addressed previously in the text, are based on continuous improvement cycles, so also here they play an important role.

Today several organizations, InfoSec professionals and other type of information sources related to the cybersecurity industry keep producing additional principles, making this a continuous development process, in (Smith, 2012) we can find several other references from other authors.

3.1.5 Cryptography

Cryptography originates in the Greek words for "*secret writing*", it is an ancient art, and can be defined as "*the science of writing in secret code*". According to (Devi.T & Pradesh, 2013), the first known document showing evidence of the use of cryptography in writing is dated from around 1900 B.C., when an Egyptian used non-standard hieroglyphs in one of his inscriptions. In Information security, cryptography is a foundation technology which ensures and supports the following properties for any data either at rest or in transit (Uttar et al., 2018) and (Dayalan, 2020):

- **Confidentiality**: It means that the content of the message remains accessible only to the sender and the intended receiver of the message.

- **Possession and control**: A mechanism that specifies who can have access to and controls the data.

- **Integrity**: It ensures that the content of a message remains the same, i.e., is not altered between the time the message is sent and the time it's received by the intended receiver. The receiver must be able to validate if the message was changed in transit.

- **Authentication:** The authentication mechanism ensures proof of identity, i.e., the receiver can verify that the sender is who he claims to be, likewise the sender can verify the identity of the receiver.

- **Non-repudiation**: A mechanism that ensures that the sender cannot deny he was the sender of a certain message.

There is some basic terminology used in cryptography that is crucial to know in order to understand the whole concept.

Cryptography Concepts

A message in its original state i.e., a readable message, is known to be in plain text, or clear text. This message is subject to the encryption process which will generate the encrypted message also known as ciphertext, i.e., a scrambled form of the plain text. The encryption can be defined as reversible conversion of the plain text into a ciphertext. Once encrypted, a message can be decrypted. This decryption process is the mechanism to convert the ciphertext back into the original plain text. There are two ways of doing decryption, the normal and expected way is to decrypt a message using its encryption key. The other one, not desirable, is to decrypt it without knowing the key. This is done by attackers by breaking the code, action also known as cracking the code. The Key mentioned here is a secret allowing encryption and decryption to be restricted to the possessors of that key. This key is what makes the encryption process secure (Uttar et al., 2018).

Types of Cryptography

- **Secret Key Cryptography**: Means the same key is used for both encryption and decryption. Examples of encryption algorithms that use secret Key Cryptography are DES, Triple DES, AES, RC5 and a few others. This mechanism is also known as Symmetric Encryption. The challenge of using this encryption is on finding a secure way to exchange the secret key between sender and receiver (Uttar et al., 2018) and (Devi.T & Pradesh, 2013).

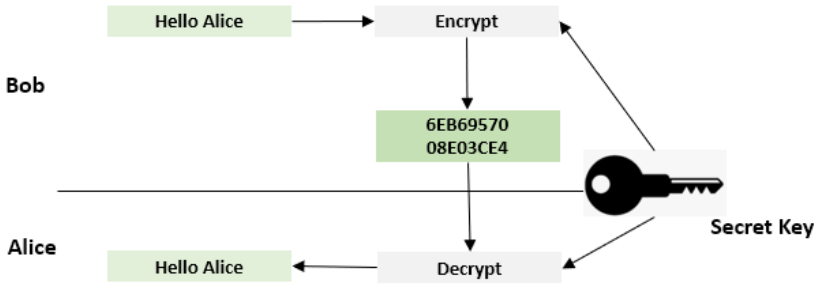

Figure 9 – Symmetric Key Cryptography

- **Public Key Cryptography**: This method uses one key for encryption and another key for decryption. As examples of public key cryptography, we have RSA, Elliptic curve, and a few other algorithms. RSA is a public key algorithm developed by Rivest, Shamir and Adleman, hence the initials RSA. They founded the company RSA Security, which has been since, a reference in the security industry. One of the keys is public and it's used to encrypt the message, the other key is kept private, only known to the receiver, which will use it to decrypt the message. This is also known as asymmetric encryption (Uttar et al., 2018) and (Devi.T & Pradesh, 2013).

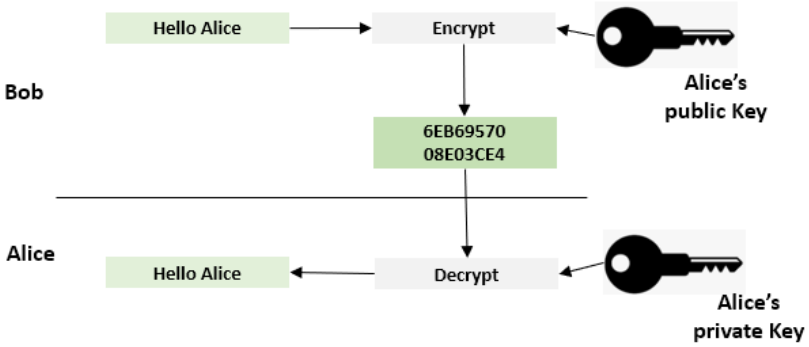

Figure 10 – Asymmetric Key Cryptography

- **Digital Signature**: Is a very important mechanism used in cryptography which ensures non-repudiation and integrity of a message. Digital signature makes use of public key cryptography to sign a digital message. A message sent to the receiver is not actually encrypted, the message is signed with the sender's private key and can be verified by anyone who has the sender's public key. This proves the sender is who he claims to be, i.e., non-repudiation. This also ensures that the message has not been altered in transit, i.e., integrity, because the digital signature mathematically links that message (message Hash) to that specific sender. If the key used is authentic, this method assures those properties remain valid (Uttar et al., 2018).

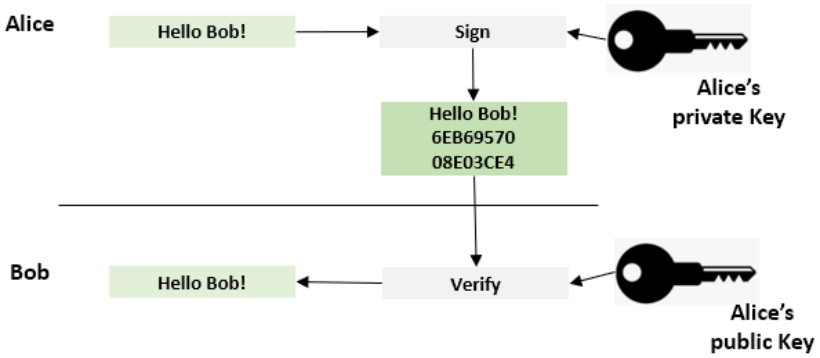

Figure 11 – Digital signature

Public key cryptography and digital signatures rely on what's called a public key infrastructure (PKI). PKI is an infrastructure in which one or more entities, known as certificate authorities, certify

ownership of public/private key pairs. This implies that the PKI system is trusted by all systems involved in a communication (Uttar et al., 2018).

- **Hash Functions**: Uses a mathematical algorithm to irreversibly "encrypt" a message. This algorithm generates a message summary, or digest, to confirm the identity of a specific message and to confirm that there have not been changes in its content. This is a one way only process, meaning there is no way to decrypt this information. The same message always provides the same hash, but the hash itself cannot be used to determine the contents of the message. Hash functions are used essentially to confirm message integrity (Uttar et al., 2018) and (Devi.T & Pradesh, 2013).

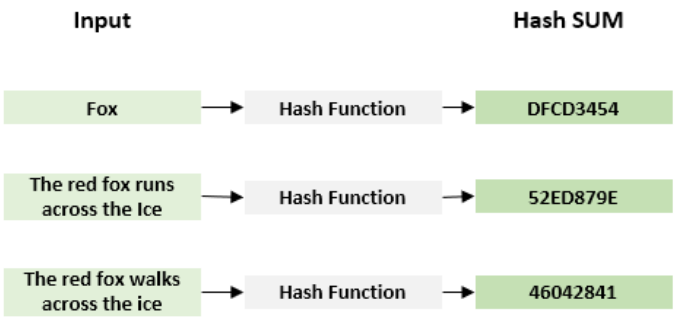

Figure 12 – Hash function

Secret Key vs Public Key Cryptography

A comparison study performed by (Chandra et al., 2014), help us identify the main aspects that differentiate these two main types of cryptography. For the sake of this comparison, one is interested in describing the main advantages and disadvantages of each mechanism.
In symmetric cryptography, encryption and decryption of information is normally easier because the key is already known by both parties, also, the encryption process does not involve too much complicated computations. The encryption itself is very secure, especially if a key with 256 bits is used, which in this case, an attacker can take several years before being able to guess the secret key. The main disadvantage of symmetric cryptography is the challenge in finding a secure way to exchange the key between the sender and the receiver. This secure way of exchanging the secret key must be ensured so we can avoid it being intercepted allowing an attacker to tamper the information. If a secure key exchange cannot be assured the usage of symmetric encryption is unreliable.
Asymmetric Cryptography has the advantage of being more convenient to use, as the problem of distributing the key doesn't apply. Everyone can use their own private key to decrypt a message that's been encrypted with a public key. The fact that it provides authentication, using digital signatures, and non-repudiation, is a big advantage also. It's much easier to identify if a message has been tampered. However asymmetric encryption is slower to decrypt, which becomes worse when there are large messages that need to be decrypted. With this it also comes the need for more computational power compared to the case where we use a single key. The individual private key which is an advantage if we look at the key exchange challenge, can also be a disadvantage in the

case the key is lost, which may lead to an attacker to gain access to all the information of a certain receiver.

3.1.6 Threats

Previously we have seen that there are several definitions for cybersecurity, one of the most known is the NIST definition "*The ability to protect or defend the use of cyberspace from cyber-attacks*" (NIST, 2013). The purpose of this work is to build a framework for cyber defence using AI, so while one is conducting the study on the cybersecurity discipline, it's of utmost importance to understand what type of threats are out there, and with it, what are the type of attacks used. Knowing the existing threats and how they are materialized through a certain type of attack, one can then talk about defence.

The existence of a threat implies that there is an attacker of some sort who's willing to inflict damage. These attackers have motives, like any criminal in the physical world. As an example, there are corporate spies that conduct cyber-attacks on some organizations of their interest to obtain business secrets. Other type of criminals may be seeking financial gain. Some attackers, just want recognition for their acts, they seek glory. The most dangerous and advanced attackers are nation states who normally seek some type of political advantage against their counterpart typically to increase their power. Nation states attackers are very well funded so they can conduct very sophisticated attacks that require good financial resources (Gordijn et al., 2020).

These well-funded or highly skilled attackers supported by organizations or nation states can be seen as 'professional' attackers, but besides these, there are also attackers who can be considered hobbyists. For example, the term "script kiddies" refers to attackers who are not very skilled and are only able to use "ready-to-run" tools in their attacks. There are also what's called "hacktivists" who basically perform attacks to support a cause and generate publicity for that cause. Moreover, there are the so called "rogue hackers" who mostly attack systems out of curiosity, with no special purpose, and not supported by anyone (Gordijn et al., 2020).

Some hackers only perform attacks for personal gain. Other type are the ones who just want to make fun of their victims by taking down their websites, just so they can brag about their capabilities to their peer hackers. Additionally, they may sell sensitive data on the dark web. The term 'black hats' is used for attackers with malicious motives. On the opposite spectrum, there are the 'white-hat hackers' whose interest is to improve overall security. The objective is to report all discovered vulnerabilities back into the organization. These can also be called "Ethical Hackers" (Gordijn et al., 2020).

Security threats are not only generated by outside attackers. There are also some crimes that originates internally within the organizations. A disillusioned employee or espionage agent can find ways to get access to company sensitive data. Data can be altered, stolen, put on the web and so on. A contractor with more access than he should, can easily access certain data which he will then be able to sell to a competitor. These insider criminals can also use their privileged position inside an organization to drop a piece of malware into a computer and infect the organization's network. Attacking a high-profile target by using a specific vendor to deliver a piece of malware is something that has occurred recently with some frequency, and it's called a supply-chain attack. This type of attack is quite powerful and very hard to detect (Gordijn et al., 2020).

Not implementing the appropriate security policies by using some of the standards and frameworks one has described previously in this text, can also contribute to the existence of additional threats. Threats originating for this lack of practise may be considered as originating within the organization. Similarly, not having cybersecurity professionals with the right skills fall into the same logic.

Types of Threats and Attacks

- **Cyber Theft:** This is the most common cyber-attack committed in cyberspace. It basically involves using some type of method to steal information or assets. Using a malicious script to break and gain access to the systems or network security without user knowledge or consent, for tampering or stealing critical data, is a type of cyber theft. Stealing credit card information, social security numbers, medical information, any type of personal identifying information is also cyber theft. Identity theft is also a very common crime of this type (Razzaq et al., 2013).

- **Cyber Vandalism:** Damaging data instead of stealing it for a specific purpose is called cyber vandalism. This normally results in access to data and systems being disrupted or stopped. It prevents the authorized users from accessing the information contained in the systems. This type of cybercrime can be analogous to a time bomb because it can be set to trigger a certain action at a specified time, to damage the target system. Cyber vandalism often disseminates harmful software which does irreparable damage to computer or network systems, deliberately entering malicious code like viruses, worms or ransomware into a network to monitor, disrupt, take down, or perform any other malicious action (Razzaq et al., 2013).

- **Web Jacking:** Web jacking is the act of forcing the control of a web server through gaining full access to the web site of another organization. After gaining control, hackers often manipulate the information on the site causing the intended damage (Razzaq et al., 2013). Sometimes they may even demand a ransom to give back the control to the owner.

- **Software Piracy**: Software Piracy is the distribution of illegal and unauthorized copies of software that is counterfeiting. This is considered illegal digital broadcasting. Unauthorized download of software, even if it is an original version is also an act of software piracy (Razzaq et al., 2013).

- **Industrial Espionage**: A certain business using spies to illegally monitor the systems or network traffic of their business competitors is an act of industrial espionage. Normally they are looking for Information of future products, marketing strategies, or financial information. The objective is usually to gain some type of competitive advantage (Razzaq et al., 2013).

- **Cyber Terrorism**: There are well organized and well-funded terrorist's networks that can conduct very sophisticated attacks, not dissimilar to the capability of some nation states. These normally have political motivations and often include violence against civilians, ethnic minorities or some other group that may be of political interest (Razzaq et al., 2013).

- **Wi-Fi Jacking:** Is using wireless networks that are not properly secured to connect to unauthorized and private systems which are also connecting to the same networks. There are a lot of wi-fi networks that either are wide open, or are using weak authentication mechanisms which make them easy targets for hackers to highjack (Razzaq et al., 2013).

- **Cyber Assault by Threat**: Using computer and internet technologies, like social networks, email, or other communications applications to threaten persons for their lives or their families, or any act of cyberbullying. As an example, blackmail people leading them to transfer funds to some untraceable bank account following such threats (Razzaq et al., 2013).

- **Logic Bombs:** This is software that is activated after some specific event being triggered. For example a virus can act as logic bomb because it can "sleep" throughout and extended period of time and become active only on a particular date (Razzaq et al., 2013).

- **Salami Attacks:** This type of attack is usually performed by using a software that is modified to retrieve small amounts, the so-called slices, of money in a financial transaction, redirecting them to a hidden account. Since the "slices" are very small compared to the amount involved in that transaction, it easily goes unnoticed. Logic bombs can be used for these crimes, for example, by introducing them into a bank's system and setting it to deduct a small amount of money from every account occasionally. Taking 20 cents from every account at random periods of time, certainly goes unnoticed and can result in a considerable amount of stolen money (Razzaq et al., 2013).

Besides the important examples described above, currently there are a few other sets of threats and attacks that are very in trend, such as:

- **Denial of Service**: A denial of service attack (DoS) or distributed denial of service attack (DDoS) is a type of action that attempts to make a computer resource, or any system whatsoever, unavailable to anyone who neds to use it. The systems of the victim are inundated with more requests than it can handle affecting its performance and ultimately taking it down. The Ping of Death or POD is a good example of a DoS attack. It works by sending a malformed ping (ICMP message) to a server or any other computer. The ping is so big in size, compared to a normal ICMP message, that the IP (internet protocol) packet can't handle it, resulting in a buffer overflow. Due to this, a system can become vulnerable and potentially allow the execution of malicious code. Another example is a ping flood, which sends thousands of ICMP echo requests. When an ICMP echo request is sent to a destination, the receiver is supposed to reply with an equal number of ICMP echo replies. The thousands of ICMP messages sent when a ping flood attack is conducted is so high that the receiver can't handle with it, its unable to respond, and often it crashes (Yadav, 2020).

- **DNS Cache Poisoning Attacks**: DNS, or Domain Name System, works by getting information about IP addresses from an FQDN (fully qualified domain name) to make it easier for a person to find a website or any other type on system on a network. A DNS cache poisoning attack works by hacking a DNS server introducing a fake DNS record that will point victim to a compromised server. Once the victim connects to the compromised server it can get infected

with some type of virus, worm or other malicious software opening a channel for either control, access to data or others (Yadav, 2020).

- **SQL Injection attacks:** This is an attack in which a malicious code is inserted into a website using Structured Query Language (SQL). This makes use of a web security vulnerability that allows an attacker to interfere with the queries that an application makes to its database. Attackers can then run SQL commands that allow them to delete website databases, copy data, modify it, or run other malicious commands (Yadav, 2020).

- **Cross site scripting attacks:** This is a type of injection attack where the attacker inserts malicious code to exploit the user of a web site. These attacks are used to highjack a user session. It can be performed just by adding a malicious script in a website, which a user will run in his web browser without knowing it once he accesses the site. The script could then do malicious things like steal a victim's cookies, have access to saved log in credentials and other information. It can also modify sensitive data which is useful for a victim (Yadav, 2020).

- **Password attacks:** User passwords normally are not as secure as they should be, so hackers perform password attacks as one of the most common ways to gain access to a user account. Software like password crackers is often used to try to guess a user's password. One of the methods used is called a brute force attack, which consists in continuously trying different combinations of characters until it succeeds. Since this attack requires testing a lot of combinations of characters, its usually time consuming. Another common method is called a dictionary attack. A dictionary attack doesn't test out brute force combinations of characters. Instead, it tries words that are known to be most frequently used in passwords by the users (Yadav, 2020).

- **Rogue Access Point Attacks:** Installing somewhere in a building a wireless access point without the knowledge of the network administrators, may be seen as something irrelevant but can be critical from a security point of view. This access point is certainly bypassing any management rules that impose some security controls on the network devices, and with that it can be providing access to the network to any unauthorized person that simply connects his computer to that hotspot. An attacker can easily use this channel to gain access to the network, even being outside of the building, and initiate any malicious action from there (Yadav, 2020).

- **Evil Twin Attack**: This one is technically similar to the rogue attack, however there is a difference, where the objective of this type of attack is to get the victim to connect to a network that is identical to his own network. The victim won't notice because he's connected to an identical infrastructure, but this network, the evil twin, is fully controlled by the attacker which is then able to monitor all the victim's traffic (Yadav, 2020).

- **Known Vulnerabilities:** Not performing regular updates either on systems hardware, or on software, leads to potential security breaches. Patching regularly is one of the most important activities in security, because it addresses vulnerabilities that may not be known by users or system administrators, but certainly known by the attackers. Attackers always

look to use systems that have these vulnerabilities to initiate their attacks. A system administrator should always conduct diligent efforts to maintain systems updated to the latest patches instead of being worried that the software stops working because of implementing this change. The cost of being attacked due to not addressing a vulnerability can be very high (Yadav, 2020).

- **Social Engineering**: Is the act of persuading a person to either provide any type of information or perform some type of action to be able to later exploit the individual or the organization the individual works for. This is sometimes "*one of the simplest ways to collect information about a victim or other target by exploiting the human weakness that exists in any organization*"(Conteh & Schmick, 2016)

- **Phishing:** This is a type of social engineering attack. Phishing scams are used by criminals as an attempt to obtain personal information, like credit cards numbers, social security numbers, passport numbers, bank account details and so many other types of personal identifiable information. In a way is also a kind of cyber theft, like described previously. These scams, or attacks, are normally done by sending links in text messages or emails, trying to lead the victim to click on it, and then getting them to insert their personal data on a fake web site. These web sites look legitimate and often are very good copies of reliable organizations websites, like banks. The user thinks he's introducing his data to access a bank application for instance, but in fact is handing over the information to criminals (Conteh & Schmick, 2016).

- **Email Scams:** Another type of social engineering attack. Fraudulent emails from people who are supposed relatives or supposedly representing family members, that lead relatives or friends to send money to hidden bank accounts for fake reasons (Conteh & Schmick, 2016).

Malicious Code - Threat Model

When we talk about malicious code, we are referring to what is commonly known as malware. This type of threat is very significant today for any systems and network security. Malware is a piece of software that is included or inserted into a system without the knowledge or authorization of its owners and administrators. Its purpose is to compromise the integrity, confidentiality and availability of the systems data, applications or even hardware. It can affect not only software but also hardware and firmware of a computer system. Malware can have multiple classifications depending on its functions and attributes. Examples are virus, worms, trojan horses, spyware. Inserting a type of malware into a system is normally done by exploiting some type of vulnerability. Alternatively, when no vulnerability is found, malware can also be inserted by convincing a victim to execute it, often without them knowing what they are really doing (Gordijn et al., 2020).
The use of malicious code can be considered to follow a specific threat model. A threat actor which can be an individual, an organization or even a nation state, accesses a resource of interest, through a physical or logical path, to perform the action of executing a type of malicious code, with the purpose of getting a specific outcome of that action. The outcome can be to retrieve intelligence information, perform some surveillance or reconnaissance of the victim organization, or it can be to disrupt a systems operation, destroy the assets, achieve publicity for his cause or eventually generate

negative publicity for his victim (Bosworth et al., 2014). This model can be depicted in the following way:

Figure 13 – Malware Threat Model

The execution of the malicious code, results in the delivery of its payload. The payload is the function or action inserted into the system, e.g., malicious logic, remote access or remote-control software (Bosworth et al., 2014).
The access used to deliver the payload is called the attack vector, i.e., the agent (avenue of access), a physical access (people who can enter the premises), network access (web server, client computer, email attachment, phishing) (Bosworth et al., 2014). The actors mentioned here can be of two types, structured threats, or unstructured threats. Structured threats are normally nation states, corporate criminals, and organized crime. Unstructured threats are usually rogue actors, e.g., individuals and script kiddies (Bosworth et al., 2014).

Types of Malwares

Malware comes in various types, and each type has specific characteristics leading to different forms of executing an attack, and different ways to defend and prevent from such attacks. Next, one describes the several types of malwares.
Viruses are a very well-known type of malware, they exist in different types, for example, logic bombs which is any malicious code, replicating or not, that delivers a payload as a result of a logic test. Time bombs are a subset of logic bombs which are configured to set off on a specific date and time. Cross-site scripting malware are another type of viruses which consists of code that exploits flaws in web servers and client code. Polymorphic viruses intend to signature-based antivirus tools and modify themselves at the time of replication. They are able to encrypt code and have a self-decrypting capability (Bosworth et al., 2014).
Worms are a different type of malware, often compared to viruses, but with different characteristics. While viruses integrate into host code and replicate upon execution of such infected code, worms are free standing code which replicate via networks. Email is a common vector to deliver worms. Some worms have viral properties, in the sense that they can integrate themselves into email messages and convert them into executable files concealed by using a different type of file extension (Bosworth et al., 2014).
Spyware and Adware, another type of malicious code, is software that collects information about users without their permission. Monitoring the usage of licensed programs or tracking and reporting of web usage is often done using spyware. Click bait adds, or unwanted adds served to a user on a web page is a type of adware. Click bait often can lead to a fraudulent page, which can steal user information (Bosworth et al., 2014).
Rootkits are programs that allow covert access to the system after installation. Once access is achieved, they can compromise applications, operating system Kernel, even the hardware or hypervisor levels (Bosworth et al., 2014).

Botnets are automated processes running on the internet. These are configured to carry out specific tasks in an automated way, like web spidering (collecting files from web, e.g., google engine bots), monitoring conversations on chat channels (e.g., for suppression of profanity or for automated responses to questions) (Bosworth et al., 2014).

How to deliver and propagate malware (deliver the payload)

Some social engineering techniques can be used to convince a victim to install malware without his knowledge. A real example is by sending an email with an attachment inviting a user to download and run the attached file, this will install the malware leaving it to perform the intended damage (Gordijn et al., 2020). Another common technique used to deploy malware is to trick a user to visit a website that is controlled by an attacker. This technique is called "drive by download", and the way it works is that the website is configured specifically to exploit a vulnerability in the user browser. This vulnerability will allow the attacker to force the browser to execute the malicious code. Often this type of malicious code is inserted into the adds that a user sees in a website. Any visitor of that web site that is not fully updated with latest security patches will be vulnerable and subject to become infected while just browsing it. Using adds to deploy malware is called "malvertising" (Gordijn et al., 2020).

Another example is the "water holing" attack, this works if a victim visits a website that hosts updates for the software the victim is using. The attacker will insert the intended malware on this compromised site, and once the victim connects to update its software or firmware, it will download the malware consequently infecting his own system. Once infected the attackers can carry out a widespread attack from it (Gordijn et al., 2020). In June 2017 a group of Russian hackers inserted a malware called "Not Petya" into the update server of an application called "MeDoc" which is used by every entity that fills taxes in Ukraine. This malware has been downloaded by every company that used that software to submit their taxes in Ukraine, and it encrypted the computers with the sole purpose of causing severe disruption. This attack has been probably one of the most devastating in history, and besides the huge technical implications, it had an economic impact of over 10 billion dollars in the world economy (Moura & Trilho, 2021). This is an example of a "water holing" attack.

Keyloggers which are tools used to exfiltrate username and passwords from user's computers are also a very common type of payload (Gordijn et al., 2020).

The use of botnets that are configured to performed actions like sending a huge amount of spam emails, or DDoS attacks are another example of a payload (Gordijn et al., 2020).

Detection and prevention of malicious code attacks

Defence in depth is a strategy that can be used to deal with malicious codes attacks. Implementing defence in depth means that we use multiple concurrent strategies: operational controls, human controls, and technical controls. Operation controls consists of written policies and procedures like who can install programs, acceptable usage policies for internet access and emails usage for example, how to respond to a suspicious attack, employment policies and other relevant procedures (Bosworth et al., 2014).

Human controls are related to any action that involves training the users, so they are knowledgeable on how to behave when it comes to certain attacks. Providing them training on malware policies and procedures, advise them of current threats like being alerted for phishing scams, not ignoring anti-virus pop-ups, being conscious of suspicious emails (Bosworth et al., 2014).

Technical controls are related with the implementation of technology that is specific to prevent attacks, like anti-virus systems, host configuration controls and security, network based security controls, network monitoring (Bosworth et al., 2014). Anti-virus systems exist of two types, they can be both host-based or network-based. Good practises are to use systems from different vendors to increase protection and make sure they are kept updated daily. When it comes to viruses on emails, there are specific systems that are optimized to look into phishing emails, spam and other types of fraud (Bosworth et al., 2014), these type of appliances are network based controls.
Automatic updates, i.e., patching of operating systems and software, removing unnecessary software, apply tight security configurations on browsers, are examples of good host security controls (Bosworth et al., 2014). The usage of web proxies to secure web browsing is also a very good security practise when it comes to prevent attacks from malicious code. Web proxies are also a type of network control for viruses.
Other type of network-based security controls is the configuration of a layered network architecture to interfere with the propagation of malware if it gets inside. Routers with proper access control lists (ACLs) to segregate layer 3 communication, firewalls to block either specific TCP and UDP ports as well as some type of applications and suspicious traffic, implementation of virtual local area networks (VLANs) to segregate networks are security controls that should be implemented (Bosworth et al., 2014). Network monitoring is also a type of network-based security control, which uses some tools to monitor and aggregate data from the several systems like devices, servers and host logs, intrusion detection alerts, and monitoring network data flow. Keeping an historical database of these event logs allows to set behaviour baselines and look for statistics which serve as a mechanism to help detecting anomalies when they occur (Bosworth et al., 2014).

3.1.7 Threat Detection Technologies

In the previous section one has addressed the detection and prevention topic mostly focused on the malicious code perspective, and here one will focus mainly on technologies that exist in the industry that are used not only for malicious code detection but also other types of threats.

Cloud access and security brokers (CASB) - Are devices that helps detect and protect cloud-based applications from being accessed in an unauthorized way. These devices detect patterns related to cloud applications access, blocking them when these patterns are suspicious. Often CASB are used in conjunction with secure web gateways (SWG), a solution that filters unwanted software/malware from user-initiated Internet traffic and enforces corporate and regulatory policy compliance. These gateways include URL filtering, malicious code detection and filtering, application controls for some web-based applications, and sometimes also some data leak prevention features (Yadav, 2020).

Endpoint Security – Are technologies designed to collect endpoint data (users' computers) and perform real time monitoring of all activity in the endpoints. This continuous monitoring analyses patterns helping detecting threats. Anti-virus, host-based intrusion prevention (HIPS), host-based intrusion control (HIC), file integrity checkers (FIC), endpoint detection and response functions (EDR) are the main security endpoint technologies. EDR are systems designed to restore data immediately in the event where encryption starts due to a ransomware attack (malware infection) (Yadav, 2020).

Intrusion detection systems (IDS) - These are devices designed to detect threats by monitoring network traffic, easily detecting malicious activity that is taking place in the network. They are normally installed in the perimeter network but can also be installed in internal networks to detect attacks. Normally they are used in conjunction with intrusion protection systems (IPS), which are designed not only to detect but also to prevent the attacks. These technologies provide detection via several methods, for example, signatures, protocol anomaly detection, various methods of analytics, behavioral monitoring and heuristics, advanced threat defence (ATD) integration, and threat intelligence (TI) to "*uncover unwanted and/or malicious traffic and report or take action on it*" (Yadav, 2020).

Network Firewalls – Are devices used to limit access to a network or a segment of a network. It works by inspecting attributes of individual network packets and analyzing them in different ways. The inspection of each packet coming into the network or leaving the network is a critical safeguard. Normally firewall have rules that act in a source to destination logic, meaning that packets originating from a certain source network to a destination network will be treated according to these rules. The rules can be "allow" rules or "deny" rules, which are based also on the TCP or UDP ports used. Packets that hit a deny rule are dropped and cannot reach the destination, packets that hit an "allow" rule are delivered. Network firewalls are of two types, stateless and stateful. Stateless firewalls only detect invalid packets, and stateful firewalls are capable of detect and also drop packets that are considered malicious (Yadav, 2020).

Network Sandbox - Sandboxing is a technique for detecting malware and targeted attacks. Network sandboxes monitor network traffic for suspicious objects and automatically submit them to the sandbox environment, where they are analyzed and assigned malware probability scores and severity ratings (Gartner, 2021b).

Secure E-mail Gateway (SEG) - Email is the most used channel for both opportunistic and targeted attacks, as well as a significant point of egress for sensitive content. Once installed, malware can obtain a foothold on the network and pivot to other endpoints, eventually finding their way to protected resources like databases, file shares and sensitive emails. A Secure Email Gateway (SEG) is implemented so that all emails are analysed and determined to be safe before being delivered to user's mailboxes (Bosworth et al., 2014).

Data Loss Prevention (DLP) – This is a solution that looks at certain textual contents in documents such as strings and numbers that look like a credit card number, information that contains proprietary information, presentations labelled as confidential, encrypted files sent by contractors and several other types of data that can be labelled according to defined categories. DLP solutions have the functionality of creating rules to act on this type of sensitive data preventing exfiltration (Bosworth et al., 2014).

Database Firewall (DBFW)- This is a type of firewall that operates at layer 7 of the OSI model, i.e., it can be considered an application firewall and it protects databases from attacks, data loss and theft. These are also known as database audit and protection tools, and they provide comprehensive security for relational database management systems. These tools have capabilities such as data discovery and classification threat, vulnerability management, intrusion prevention, activity blocking,

identity and access management analysis (Gartner, 2021a). These are the type of tools that are used to prevent attacks like SQL injection which was described previously in this text.

Mobile Threat Detection (MTD) – These solutions are designed to protect organizations from threats on iOS and Android devices. MTD solutions ensure protection of mobile systems at the device, network and application levels (Gartner, 2018).

Security Information and Event Management (SIEM) - Are solutions which provide a centralized location to collect log files from disparate sources, correlation of data (in real-time or as background process). A SIEM solution analyses event data in real time for early detection of targeted attacks and data breaches, and to collect, store, investigate and report on log data for incident response, forensics, and regulatory compliance (Gartner, 2021c).

Threat Intelligence – Cyber Intelligence is shared information that provides ways to identify attacks, adversaries, and malware, and prevent attacks from taking place as a result. The threat intelligence systems are a type of platform where is kept information about previous attacks and threats that have caused significant damage and impact. With this available information, if a company detects a threat or attack to its systems, they can consult this information and get some insights about what kind of impact it's expected and what systems may become compromised. Insights about how to tackle the situation based may also be available (Yadav, 2020).

3.1.8 Threat Prevention

When it comes to prevention there are a set of best practices that anyone should follow. In section 3.1.6 it was addressed the prevention and detection of malicious code, and in section 3.1.7 one has talked mostly about detection technologies, but some prevention technologies have also been described.
Here one will focus mostly on actions or best practices for prevention and not much on the technologies.

Passwords- User education relative to the use of strong passwords is a crucial security practice which can avoid a lot of problems. The usage of secure passwords can not only be done with user awareness programs but also with implementing policies in the identity and access management system (IAM): forcing rules like minimum 8 characters long, a combination of small and capital letters, numbers, and symbols. These rules make it much more difficult to hack a password, especially if an attacker is using a dictionary attack. Additional password rotation rules should be implemented, like forcing a password change every 90 days, and not allowing to repeat a used password (Yadav, 2020).

Regular Backups – Any type of information that is important for a user or an organization must be included in a regular backup strategy. Either being personal information, sensitive information, credentials details, business information and so on, if not backed up are subject be lost. Not only due to a cyberattack, but also a system failure can lead to data loss. With a regular backup schedule and a well-defined backup process, data loss prevention is ensured by the organizations. When defining a

backup strategy it's important to perform regular restore procedures to ensure the backups are working, and also that we are able to restore data in case of a data loss event (Yadav, 2020).

Auditing and Logging – This has been addressed when talking about technologies like intrusion detection systems, endpoint protection, firewalls, SIEM, ensuring auditing of any even that occurs in a network is a critical threat prevention measure (Yadav, 2020).

Disk Encryption – Encrypting valuable data is also an important prevention measure. A hypothetic internal hacker that tries to steal a physical disk won't be able to do anything with the data without the decryption key. On laptops, which are devices that users carry with them everywhere, using an encryption tool like "bit locker" is also a very important security measure (Gordijn et al., 2020).

Preventing Injection Attacks – In section 3.1.6 we looked at several types of threats being one of them the SQL injection attacks. As a prevention measure for these attacks, we can use safe APIs for development. Using LIMIT or other SQL commands in the queries can help mitigate disclosure of data records in case of a SQL injection attack (Yadav, 2020).

Multi Factor Authentication – Is an authentication method that requires more than one type of proof of identity to gain access to a system. For example, it can combine a password (something you know), some type of token like an SMS code (something you have) and some biometric characteristic like a fingerprint (something you are). This prevention control is a deterrence technique, which in practical terms means that it increases the effort for an adversary trying to perform the attack, therefore contributing for the target to become less attractive from the attacker point of view (Gordijn et al., 2020).

3.1.9 Economic Relevance

Cyberattacks have today a very important impact in the world's economy. In recent years there have been reports of hundreds of cyberattacks which resulted in data leaks for millions of users and several billion dollars of costs due to these attacks. In the year of 2017 there was an estimation of more than 1500 data breaches exposing data for millions of people, 37% more that in 2016. Equifax, a big credit reporting agency in the US, revealed that a breach exposed credit card information of about 143 million people. This had an estimated cost for the company of 1 million dollars in additional services offering for their customers. Verizon, a telecommunications giant, got attacked and this breach led to the exposure of names, phone numbers and account PINs for circa 14 million customers. The economic impact is not reported though. A company that provides a platform for social learning called Edmodo, also got attacked in 2017, leading to the disclosure of data from 77 million users. This data was put on sale in the dark web, and the hacker was asking for $1000 USD for the whole database (Strain, 2018).
Also in 2017, the Not Petya attack, as already mentioned previously in this text, had a global impact of over $10 billion US Dollars (Moura & Trilho, 2021).
Moving forward to the following year, 2018, we can see that, statistics on cyberattacks got even worse. British Airways got the data of 308.000 users stolen; this data was probably credit card information. T-Mobile, another big player in the telecommunications market, suffered a 2 million user data hack. Facebook, 29 million. MyFitnessPal 150 million, and the list goes on (Salim, 2019).

According to (Morgan, 2020), during the next five years, the global cybercrime costs are expected to grow 15% every year, potentially reaching the value of $10.5 Trillion USD in 2025 compared to the cost of $3 trillion USD reported back in 2015. The prediction just for the year of 2021 is for $6 trillion USD, and to put this into perspective, this means that if cybercrime was a nation, we would be talking of the world's third largest economy, after the United States of America and China.
Looking at these numbers we can say that cybercrime will soon become more profitable than the trade of illegal drugs in the whole world (Morgan, 2020). The costs of cybercrime are related to destruction of data or its theft, money theft, loss of productivity infliction, intellectual property theft, personal and financial data theft, fraud, disruption of businesses or critical infrastructures like oil, gas and water facilities (Morgan, 2020), and even nuclear facilities as was the case of the Stuxnet attack in 2010.

3.2 ARTIFICIAL INTELLIGENCE

Nowadays computers are required to solve problems that are getting more and more complex, and in some cases its resolution is no longer dependent on increasing their computational power. Due to this, with the course of time, scientists have been trying to incorporate in computer programs and systems, knowledge and capabilities which are normally associated with the human being. This is the main objective of artificial intelligence (AI) (Costa & Simões, 2008).

3.2.1 History and Background

Knowing if a machine will be able in the future to have an intelligent behaviour compared to a human is an ancient question. However, it was not until 1950 when digital computers became more common and accessible that, following the idea from Alan Turing that machines could become intelligent, the scientific community embraced projects to build some. First computers were dedicated to scientific and military calculations, but progressively its applications extended to other areas of human activity. With the end of the second world war, people started to evaluate the possibility to use computers in other activities besides the sole purpose of military applications. At that time one of the areas which started to gain attention by the scientific community was Artificial Intelligence. In 1956, at the Dartmouth University in the North American State of the New Hampshire, took place the first scientific conference which gathered several scientists that were by then pioneers in the AI area (Oliveira, 2019).
Promising names like McCarthy and Minsky were present at that conference (Lu, 2019). The first approaches to this thematic tried to reproduce part of human thinking, which were seen at that time as the most advanced, like demonstrating theorems, planning sequences of actions and play board games (Oliveira, 2019). The objective was to find a way to make a machine think like a human being, in essence: how to make it communicate in a natural language, how to empower it to have some kind of intelligence. This gathering at the Dartmouth university is known as the origin of AI (Lu, 2019).
In the decades that followed these first experiments, a multitude of other relevant work was done, but it was between the years 1980 and 2000 that the second major development of AI took place. The Japanese government did a considerable investment to support AI research in a quest to build

what they called the 5th generation computer program. The objective of this program was to build a machine that could support human to machine dialog, do translations and recognitions. Another important development in AI took place between the years of 1993 and 2000, when IBM developed the “deep blue”, and Google developed the “AlphaGo” which was the first computer program to beat a human in a board game, in this case the world champion Li Sedol (Lu, 2019).

So, we see that in the 1990’s the horizons have widened to the point of actually building the first intelligent entities. Researchers in this decade got involved in long and hard debates to prove their approach to AI was superior to others, however going into the XXI century most of the problems that are expected to be solved with AI were still without solution (Costa & Simões, 2008).

Despite the numerous doubts, there was a consensus that the future of AI is based on the quest for solutions for the development of autonomous agents that are capable of learning, deal with uncertainty, deal with complexity and with the capability of evolving by adapting to the environments where they live. Other field of this future may be on the possibility of having not isolated agents but a society of agents that work together in cooperation to solve difficult problems (Costa & Simões, 2008).

To put it into context, an agent is an entity, that can capture information from the surrounding environment where it lives, allowing it to act on this environment, using processes that helps it define the best action to take, therefore resulting in an outcome. These processes and decisions will be more sophisticated the more complex is the task to execute or the environment where the agent is (Costa & Simões, 2008).

Agents own a set of important properties: they have a body, a location (existence in space), a set of mechanisms to understand and act on the environment, and a mechanism that supports intelligent activity (decision making). Based on these properties, agents are able to build strategies which allow them to succeed in executing the tasks or solving the problems that are imposed by the environment (external) or determined by objectives of the agent itself (internal) (Costa & Simões, 2008). Later in this text one will explore and describe the different types of agents that are used in AI.

Starting the new century, in the beginning of 2000, the main focus of scientific work on AI was on building theoretical foundations and Machine Learning (ML) methods to process and analyse sets of data. This was a period where the growth of data and internet was exponential, and we were entering the information age. Huge amounts of data were starting to be collected and treated in the cyberspace. By the year 2012 the AI field was now dealing with this data and for that it was making use of algorithms that had been through several iterations of trial-and-error approaches. Here was a point where deep learning got into play. At this point we started to see the concept of deep learning in networks appearing in the AI space (Lu, 2019).

In recent years is known to all of us how rapidly technology has been evolving, namely with the growth of internet and cyberspace and the growth of distributed and cloud-computing, in general, computer power had an exponential growth. This, together with the explosion of Big-Data, lead to the inevitability of AI becoming a hot topic today (Lu, 2019).

According to (Oliveira, 2019), it’s reasonable to expect that in a near future, most probably before the end of the XXI century, there will exist systems (agents) that are able to demonstrate intelligent behaviours that are very similar if not indistinguishable from those of a human being, when faced with a certain type of situations. This is already raising some critical questions in society like: will these systems have a conscience, will they have feelings, will they have their own plans and motivations? These are difficult questions that will have to be answered before humanity being faced

with such reality. So, the study of the AI field has been very important in the past decades, and continues to be, since there are a lot of possibilities and a lot of challenges ahead.

3.2.2 What is Artificial Intelligence

AI can be defined in a simplistic way as a discipline or field of study whose objective is to study and build artificial entities with cognitive capabilities that are similar to those of human beings (Costa & Simões, 2008). This definition varies according to literature, which is justified by the variety of point of views and the several different sources that are used by the scientific community (Costa & Simões, 2008).
(Russell & Norvig, 2021) sates that "*the field of Artificial Intelligence (AI), is concerned with not just understanding, but also building intelligent entities – machines that can compute how to act effectively and safely in a wide variety of novel situations*"
The concept is of course built on top of the idea of intelligence, which for some researchers is defined in terms of the similarity to the human performance (behaviour), whilst others prefer to define it as rationality (reasoning) (Russell & Norvig, 2021). From these two dimensions there are four possible combinations, which resulted in the four types of approaches for studying AI that researchers have focused on: acting humanly, thinking humanly, thinking rationally and acting rationally (Russell & Norvig, 2021).

Acting Humanly

A machine could be seen as acting humanly if it can pass the Turing test, which is a test designed by Alan Turing in 1950. In this test a machine would have to answer to a set of written questions posed by a human interrogator. If by reading the computer written responses, the human interrogator couldn't tell if the responses were provided by a human or a machine, the machine would pass the test. To achieve this, the machine would need some well-defined capabilities (Russell & Norvig, 2021):

- **Natural Language Processing** – This allows the machine to successfully communicate in human language
- **Knowledge Representation** – This allows the machine to keep record of what it sees or hears
- **Automated Reasoning** – The machine would be capable of answering questions and from those answers draw new conclusions
- **Machine Learning** – The machine can adapt to new circumstances, recognize, and detect new patterns and behaviours.

The four capabilities described here represents the view by Alan Turing, who believed that the emulation of the physical capabilities of a human was not necessary to prove intelligence.
However, other scientists had a different view, and in their opinion, besides these capabilities, a machine should also have some physical representations of a person, in order to be considered it has intelligence (Russell & Norvig, 2021):

- **Computer Vision** – Capability to not only view the world but also perceive it
- **Speech Recognition** - Capability to recognize a person's speech

- **Robotics** – The capability to manipulate objects and move around

To include these new variables, the scientists proposed a change to the Turing test, and it was then called the total Turing Test. What we see here, when looking at the desired capabilities of a machine, so we can consider that it has intelligence, is what in practical sense resulted in the disciplines that mainly compose AI (Russell & Norvig, 2021).

Thinking Humanly

This is considered a cognitive approach model for AI. The idea behind this approach is that if a machine thinks like a human, it must know how a human being thinks. There are three different ways that we can use to learn how human thinking works (Russell & Norvig, 2021):

- **Introspection** – This is when a human focus and register his own thoughts as he goes on doing his normal life
- **Psychological Experiments** – This is when we learn by observing how a human act and performs
- **Brain Imaging** – This is done by observing how the human brain acts.

If we can gather enough information by using these three methods, we may be able to create a computer program that is capable of expressing in theory what we have learned (Russell & Norvig, 2021).

Thinking Rationally

This approach is based on the "laws of thought" as it started in ancient Greece, when famous philosophers like Aristoteles tried to define what is "the right thinking", in a sense that the "right thinking" is no more than an irrefutable reasoning process (Russell & Norvig, 2021). An example of a logic thought is: If Bob is a man and all men are mortal, then we conclude that Bob is mortal. This type or reasoning led to the initiation of the study field of logic.
If we define a specific notation for statements about any object and the relations each object has with each other, we have defined a logical notation. Using this approach, we can then teach computers to solve any problem described in logical notation. The usage of this approach in the AI field is done by scientists that are known to follow a logicism tradition. This type of approach is not enough to result in intelligent behaviour though. For that we must add some type of rational action (Russell & Norvig, 2021).

Acting Rationally

This approach is also known as the rational agent approach. An agent, as one has briefly described previously, is an entity that acts on its environment producing some result or outcome of its actions. Agents in AI are expected to do something more intelligent than what a simple computer program does, a computer program is also producing an output as a result of its actions, so it's also a type of agent. However, the objective of an AI agent is that it can act autonomously, perceive the environment surrounding it, persist over a long period of time, adapt to change, define new goals

and try to achieve them (Russell & Norvig, 2021). If we look at some of the capabilities one has described in the previous approaches, like the ones for the logical thinking and the ones required to pass the Turing test, they all allow an agent to act rationally. An agent will make good decisions if it has the right knowledge and does a good reasoning. Other learning capabilities allow an agent to improve its ability to create a more effective behaviour especially when adapting to new circumstances. However, this rational agent approach to AI, brings additional benefits when compared to the others. This is a more general approach than the "laws of thought" of the "Think Rationally" approach because the logic reasoning used in that approach is just one of the possible ways of achieving rationality. One second addition is that this agent approach is more open and responsive to scientific development (Russell & Norvig, 2021).
These additional benefits of using the agent approach resulted in the fact that, throughout time, this approach has prevailed over the others. Using this approach, we are in better conditions to develop an agent, i.e., a machine, that does the right thing (Russell & Norvig, 2021).

Besides the four approaches one has addressed now, there is also another way at looking to the AI phenomenon. This is done by using a paradigm approach to understand AI, which is based on three types of metaphor's: The computational metaphor, the connectionist metaphor and the biological metaphor (Costa & Simões, 2008).

Computational

in this metaphor we assume that we should look at AI as a computational thing only. This principle is defended by Allen Newell and Herbert Simon, two scientists who believe that computer and human mind are part of a family of artifacts that are called physical symbols systems. These systems have the capability of processing symbolic representation of knowledge, which with the course of time will result in the generation of new symbolic structures. They define intelligence as a result of the action of processes over symbolic structures (Costa & Simões, 2008). This computational approach is materialized through programs that act upon representations of the physical world.

Connectionist

This metaphor looks at AI as a result of the interactions between a high number of elementary units of processing. This is a very generic and broad idea but is based on the concept of the relation with the human brain. In the connectionist approach the idea is that the cognitive function can be modelled as being the result of the simultaneous interaction of many identical units, which are similar to neurons densely connected. This results in the creation of artificial neurons, whose first version was proposed in 1943 by two scientists called McCulloch and Pitts. Using artificial neurons, we are able today to build artificial neural networks. The main difference between the computational approach and the connectionist approach is that the latter can't model the reality at a symbolic level (cognitive) (Costa & Simões, 2008).

Biological

This metaphor used to build intelligent agents is based on an analogy done with the way species evolve. According to Darwin evolution model, species evolve due to a process of natural selection

which promotes the survival of the species that are fitter and more capable to adapt. During reproduction, the genetic material of individuals is subject to changes determined by genetic operators. These operators promote genetic mutations in individuals that can result in a more advantageous situation leading them to become fitter and therefore having better chances of surviving. A real example that clearly illustrates this process, is the capability that some species have to camouflage, so they can avoid predators. In this approach, in order to solve a problem, we start from a group of candidate solutions which is designated as population, that we promote its evolution over time according to Darwin's principles. Each individual in this population has a quality and a merit which defines his adaptation level. The ones with better quality have a higher probability of surviving, consequently of reproducing. This artificial evolution model is a simplification of the natural evolution model proposed by Darwin (Costa & Simões, 2008).

AI has we can see, involves a wide variety of components and its usage nowadays is seen in vast areas of society. AI can be used in learning processes, for playing games, like chess on the example of IBM's "Deep Blue" or Google's "AlphaGo" we've seen earlier, but also in more complex and smarter games. Self-Driving cars that already exist today also rely on AI technology. NASA used an AI remote agent program to perform on-board autonomous planning and scheduling for the operations of a spacecraft. AI is also used in machine translation and speech recognition, Microsoft being a leading example in this space. In medicine AI is also widely used today to help doctors diagnose some conditions, especially when these diagnoses are done using medical images (Russell & Norvig, 2021). The list is extensive, so later in this text one will present a more detailed approach regarding the areas of AI utilization in today's society.
Since the objective of AI is to build these intelligent agents that one has been describing, it's relevant to provide an overview of the types of agents that exist.

3.2.3 Agents

One has started the previous section by stating that the objective of AI is to study and build artificial entities with cognitive capabilities that are similar to those of human beings. These entities that one is referring to, are called intelligent agents. One has also described the concept of agent as per (Costa & Simões, 2008). Moreover, (Franklin & Graesser, 1997) defines an agent as "*A System in an environment capable of perceiving this environment and act upon it, throughout time, with the purpose of achieving his own purposes, in order to affect what will be perceived in the future*". Another good description of an agent is provided by (Russell & Norvig, 2021), "*An Agent is anything that can be viewed as perceiving an environment through sensors and acting upon that environment through actuators*". To give a real example we can say that a human is an agent whose sensors are his eyes, ears, nose, and whose actuators are his legs, arms, voice. If we think of a machine, like a robot, its sensors might be a camera, sound detectors or proximity sensors, etc, and some motors may be its actuators (Russell & Norvig, 2021).
In AI there are five types of agents:

Reactive Agents

These are simple machines, without internal state which are limited to react to the stimulus they receive from the environment. As an example, one can provide the Braitenberg vehicle (Braitenberg,

1998). This is a system with a sensor at the front and a motor (actuator) in the back. The stronger the source captured by the sensor, the faster the motor will act (Costa & Simões, 2008). So, this agent has a basic reaction to a stimulus.

Search Agents

These types of agents are capable of solving problems. A problem can be defined by the possible combinations of its configurations, or states. One of these states is the initial state whilst others are the final states, i.e., those that fulfill the objective set by the problem to solve. The set of all states are called the search space. In order for the agent to find a solution for a problem it needs to have a set of capabilities like perceive its states and build an internal representation (model) of those states. It must be capable to act on those states, according to the rules of the problem, forcing transitions between states, and obviously it needs to be able to understand when it reaches the final state (end result). The strategy to achieve this, is by navigating through the states space searching for a path that leads to the final state. This is therefore the paradigm of search through space (Costa & Simões, 2008).

Knowledge Based Agents

These are also referred as logical agents. These agents are designed so they can form a representation of the world. Every intelligent agent lives in an environment, i.e., its world, and it's constantly trying to understand and act on this world with the purpose of achieving certain objectives. In order to successfully fulfill the tasks required to reach its objectives, the agent has to build a conceptual representation of its world, and for that it needs a mechanism of knowledge representation. This agent interacts with the environment in an intentional way based on the representations it possesses of the environment, and for that it needs to have a reasoning mechanism, which then allows it to make decisions (Costa & Simões, 2008). Humans know things because they build an understanding of the surrounding world, which in turn allows them to do things, which result from the decisions they take (reasoning process) (Russell & Norvig, 2021).

Learning Agents

Learning is the main characteristic of intelligent beings, to the point of artificial learning becoming a critical area of AI. Artificial learning has three main objectives which are: the development of computational theories for learning, the implementation of systems that have learning capability and the theoretical analysis and development of generic learning algorithms (Costa & Simões, 2008). Learning agents are the result of achieving these three objectives. An agent learns if it can improve its performance after making observations about the world. When this agent is a computer we´re talking about machine learning (Russell & Norvig, 2021).

Adaptative Agents

We know that in nature, according to Darwin's theory, the ones that evolve and adapt better have higher probability to reproduce and survive. If we bring these principles to AI, we can try to explore the possibility of building agents with capabilities to adapt. These agents are the result of evolutionary algorithms which come from four types of families: genetic algorithms, genetic programming, evolutionary strategies, and evolutionary programming. The genetic algorithms use a

population of candidate solutions for a certain problem. The individuals from that population are then selected according to their quality. Reproduction of these individuals is based on the information exchange of their progenitors in a process known as recombination. The new individuals that are generated by this process can yet be changed in a localized way through the effects of an operator called mutation (Costa & Simões, 2008). The algorithm evolves by adapting solutions to the problems it´s faced with. This in essence is the biological approach to AI that one has addressed in the previous section.

3.2.4 Areas of Artificial Intelligence

AI can be framed into several types of areas:

Language Understanding

AI can be used to understand natural language and respond to it. It basically does this by either translate a spoken language to a written format, i.e., transcriptions or live captions, or it can translate one natural language to a different one, i.e., live translation. This is achieved using mechanisms of speech recognition and understanding, semantic processing, question answering, information retrieval (Verma, 2018).

Problem Solving

Another important area in AI is the ability to formulate a particular problem with a certain representation. This is done by planning its solution knowing what information is needed and when, as well as how to obtain it to get to the result. Able to prove a theorem using plausible or inductive deduction, use interactive methods to achieve a solution, perform automatic program writing or perform heuristic searches (Verma, 2018).

Game Playing

AI can be used to play games, by accepting and understanding the set of rules of some games. Examples are chess, checkers, go, and few others. The games rules are translated into a structure allowing the machine to use its problem-solving and learning abilities to reach the end goal with a good level of performance (Verma, 2018). Previously we've seen examples of IBM Deep Blue and Google's Go which were able to defeat a human.

Visual Observation

Pattern recognition and image processing falls in the AI area of visual observation. This capability allows a machine to analyze and perceive a scenario by relating what it sees with an internal model which represents its knowledge of the world or environment it lives in (Verma, 2018).

Learning and adaptation

The machine has the capability to learn and adapt to new circumstances based on its previous experiences. This learning contributes to its development and assimilation of general rules about the world. Its capable of creating new concept and paradigms (K. Hussain, 2018).

Reasoning

Is the ability to demonstrate logical deductions. It can answer questions and draw new conclusions based on what it sees, what it learned previously and what it perceives from the surrounding. The machine can get to a point where it's able to make decisions based on some kind of acquired logic (K. Hussain, 2018).

3.2.5 Artificial Intelligence Branches

Artificial intelligence technologies can be implemented through several different techniques. These techniques fall into different categories which are also called AI branches. These branches are based on the different machine's capacities of learning, use past experiences to predict future decisions, perform automated functions, vision functions, planning and also processing natural language (Pedamkar, 2020). One shall analyse each of these branches in more detail.

Artificial Neural Networks

The concept of artificial neural networks is based on the biological metaphor of an intelligent agent as described in section 3.2.2. This is inspired in a model proposed by the two Nobel laureates scientists, David H. Hubel & Torsten Wiesel, who discovered the simple cells and complex cells in the primary visual cortex of the human brain (Ongsulee, 2017). The computer systems that make an artificial neural network are therefore based on the analogy with a biological neural network (Ongsulee, 2017).
The human brain is capable of performing massive parallel computation which enables it to perform complex cognitive, perceptual, control and recognition tasks in a very successful way. This capability is what motivates the artificial neural network systems, or ANN's (Fine, 2016). The brain is made of more than 10 billion interconnected cells that are called neurons. These neurons have the capability of receiving, processing, and transmitting information to other neurons using biomechanical reactions. Based on this model, ANN's have been developed using mathematical models that mimic this biological model of the brain. Like a neuron in the human brain is the basic element of the biological neural network, in an artificial system, the basic processing element is called an artificial neuron, or a node (Abraham, 2005). A neuron in an ANN is a mathematical function whose main purpose is to gather and classify information and transmit it to the next node of the network. The effect or reaction to this information, the neuron impulse, is computed at the weighted sum of all input signals. This results in a learning capability of the neural network (Abraham, 2005). In a simple way we may define a neural network as a set of algorithms that are designed and used to discover relationships between sets of data using a process that simulates the way a human brain operates (Fine, 2016).

Artificial Immune Systems

Like the artificial neural networks, the concept of artificial immune systems (AIS) is also based on the biological metaphor of artificial intelligence. The type of computation used in this AI branch is inspired, and relates to, many aspects of the natural immune systems (Read et al., 2012). Algorithms developed for AIS are created with two main objectives: One is to create solutions for engineering problems using concepts that are based on the way that natural immune systems work. Secondly, to

create models and simulations to study immune systems related theories (Read et al., 2012). This idea comes from the fact that there are properties of the natural immune systems that from the point of view of engineering are considered very interesting. The way immune cells of our bodies are organized, how this cell system operates in a distributed way across our bodies, how it detects anomalies when they occur enabling the organism to recognise pathogens, leading to an immune response, makes the whole concept interesting from the engineering point of view to the point where researchers and scientists on the AI area trying to replicate it in an artificial way. So, AIS's are built trying to reproduce these properties of natural immune systems (Read et al., 2012), being defined by (Castro & Timmis, 2002) as: "*Adaptive systems, inspired by theoretical immunology and observed immune function, principles and models, which are applied to problem solving.*"

Machine Learning

This is the branch of AI that is most known and the most demanding field. Machine learning (ML) is a technique that provides computers the "*ability to learn without being explicitly programmed*" (Ongsulee, 2017). Normal computer programs are static, i.e., the code doesn't change, and the computers will only execute what's in the code. Machine learning is based on the development of algorithms that learn from the data and make predictions based on that data. This also allows them to build a model based on its input data and make decisions based on what it learns from the data (Ongsulee, 2017). There is an evolution, a learning, as an outcome of this process. In applications where static programming is difficult, or even unreliable to apply, the utilization of machine learning is the best option. Tasks like securing email systems, i.e., the detection of SPAM and other malicious emails, intrusion detection systems and other types of malicious attacks are areas where machine learning is widely used. Machine learning is also widely used in the area of data analytics (Ongsulee, 2017). There are several different methods of machine learning:

- **Supervised Learning**

  In this type of learning the agents, i.e., algorithms, are trained by a supervisor, using labels associated with training examples, for instance, providing an input for which the desired output is known. Using this method the learning algorithms create models from this training data and then use these models to classify other data that is not labeled (Cord & Delany, 2008). As an example, we can provide as inputs different cameras images, and for each image we say if it's a traffic light, a cyclist, a car and so on. This classification of the image is the label. The algorithm will then learn a new function, and when presented with a different image he can predict the appropriate label for that new image, based on the labels that it's been provided before (Russell & Norvig, 2021).

- **Unsupervised Learning**

  In this type of learning, agents use data that has no historical labels. The program doesn't know what's the answer for the input data, it must be able to find out by itself what's being presented as an input. The purpose of this, is for the system to try to find some structure in the data (Ongsulee, 2017). Clustering is a common unsupervised learning activity. This comprises of detecting useful information from lots of input examples. If we feed into a machine vision system millions of pictures, the system can try to identify a pattern based on

similar images, which could be identified as some known object (Russell & Norvig, 2021). This technique is typical used when treating transactional data (Ongsulee, 2017).

- **Semi-Supervised Learning**

  This technique is a mix of supervised and unsupervised learning as it uses unlabeled and labeled data for training. Most commonly, what's used is a large amount of unlabeled data and a small amount of labeled data. This is because is cheaper to get unlabeled data and it's also easier. Face recognition is an example where semi-supervised learning is used (Ongsulee, 2017).

- **Reinforcement Learning**

  With this method the agent learns by reinforcement, i.e., by getting either a reward or a punishment for the actions it took. In playing a game for example, we can provide a reward if the program moves all the pieces in a certain way or provide a punishment if it moves in a different way. The agent has the objective of maximize the rewards and minimize the punishments, because the reward is defined as something good and the punishment as something bad. The agent notices or decides which actions it took prior to the reward and which actions it took prior to the punishment, and this way it can learn what it needs to change in order to get more rewards and less punishments (Gudwin, 1999). This is basically a trial-and-error approach, with the objective of learning the best policy to achieve better results. Reinforcement learning is commonly used in gaming, robotics or navigation systems (Ongsulee, 2017).

- **Deep Learning**

  This method of ML is used in the artificial neural networks one has described previously, i.e., the neural networks are trained using deep learning techniques. Using deep learning, an agent improves and becomes better by itself. In deep learning the use of the word "deep" is related to the fact that the electronic circuits of these systems are setup using many layers of processing which results in a computation process where the path between the input of data and the output is done in many steps (Russell & Norvig, 2021). These several layers of processing are called "*deep nets*" (Ongsulee, 2017). Each processing layer in this network uses the output data of the previous layer as it's input (Ongsulee, 2017). Deep learning systems are mostly used in machine vision technologies, speech and audio recognition, and also natural language processing, where over the years they have shown evidence of being better and more advantageous to use when compared to the other learning methods described before (Russell & Norvig, 2021).

Machine Vision

This AI branch uses specific technology that is capable of electronically perceive and understand an image. Image capturing is done by utilizing digital cameras that are built into the machines, or systems. The cameras are the detector devices embedded in the machines that can sense the electromagnetic spectrum in its range (Flores-fuentes et al., 2014), it captures the analog image and converts it into digital data. These imaging capture techniques are widely used in robot navigation, health applications like medical images analysis, pattern recognition applications, optical character

recognition, amongst others. This computerized vision enabled the machines to duplicate the abilities of the human vision (Flores-fuentes et al., 2014).

Automation and Robotics

Robots are machines that can use AI and combine all the areas mentioned in the previous section, i.e., a robot usually is able to do language understanding activities, problem solving, play games, perform patterns recognition, and image processing. The combination of all this results in the ability to have motor functions, allowing the machine to move itself through space, move other objects, and perform some autonomous tasks (K. Hussain, 2018). There are numerous real-life examples of automation and robotics usage like in military, transportation, industrial automation, self-driving cars, even in our households.

Fuzzy Logic

The term fuzzy means something that isn't very clear, or it's vague. Often, we are faced with situations where is not easy, most times even very difficult to understand if a certain condition is true or false. Fuzzy logic, in a simplistic way, is a mechanism or technique that, by measuring the probabilities of certain hypothesis being correct, tries to modify and represent uncertain information into some level of certainty. Fuzzy logic is used to provide some reasoning into uncertain concepts. By using machine learning techniques to implement fuzzy logic systems it's possible to mimic human thinking in a logical way, incorporating decisions that fall between the true or false. Whenever we have a situation that cannot be simply defined as a binary true or false, we can use fuzzy logic systems which are able to introduce intermediate levels of categorization like partially true or partially false (Pedamkar, 2020).
Fuzzy logic techniques are often used with neural networks techniques to solve engineering problems where traditional approaches are unable to provide a precise solution. When combined this way, they are called Neuro-Fuzzy Systems (Vieira et al., 2004).
(Klir & Yuan, 1995) states that when using fuzzy systems, or "fuzzification" methods, there are gains of "*higher expression power*", an "*enhanced capability to model the real world*". This is a methodology that, by exploiting the "*tolerance for imprecision*" results in the achievement of engineering solutions that are more robust and with lower costs.

Natural Language Processing

This branch of AI referred as the field of NLP (Natural Language Processing) is comprised of several subjects whose focus is on the computational processing and understanding of human languages. This field is also commonly referred as computational linguistics, and the objective of the work in this area is to build computational models that can understand the human language (Otter et al., 2021). This involves speech to text conversion, documenting speech like transcripts or translations. Converting a text to audio is also part of the activities in the field of NLP. Examples of utilization of NLP techniques are in IVR (interactive Voice Response) systems used in call centers, where an automated attendant captures the voice of a human and takes actions based on the interpretation of what is being said. Translation tools like google translate, or spelling check tools in word processors are also examples of NLP utilization (Pedamkar, 2020).

NLP can be divided in core areas and applications. Core areas focus mainly on language modelling, which means trying to quantify associations between words, dealing with the segmentation of components of the words and identify the real parts of a speech. The applications area is the part that focus on extracting useful information from texts, translation of texts between different languages, summarization of texts, classification of documents, amongst others (Otter et al., 2021). NLP is a data-driven field which uses machine learning together with statistical and probabilistic computational methods. The techniques used in the past were different than the ones used today, although also based on machine learning, it used older methods like decision trees, random forests, or support vector machines. Today, the machine learning methods used in NLP are completely based on neural models (Otter et al., 2021).

3.2.6 Applications of Artificial Intelligence

The utilization of AI in today's life is touching a very wide spectrum of society, and it will continue to grow as technology and AI techniques evolve. The following examples demonstrate the relevance and importance that AI usage is gaining in our society.

Education

In the field of education, AI can be used to address some relevant aspects like automating basic activities like grading, e.g., multiple choice exams, adapt educational software to student's needs, identify improving points for courses. AI driven programs can also be used to provide helpful feedback to teachers and students, e.g., they can monitor students' progress and if required provide an alert to the teacher regarding his performance. The way students learn can also be changed with AI, eventually some teachers may be replaced by an AI program on certain courses (Verma, 2018).

Finance

The algorithms for stock analysis and stock exchange are leveraging AI today. Finance market analysis, finance information mining, portfolio management and administration are examples of AI applications in the finance world (K. Hussain, 2018).

Industry

The AI based robotics systems which were described in the previous section are the real example of how this industry leverages AI technologies. Robots are nowadays essential in most industries and in some cases, they even replace human work. This of course brings some ethical challenges, but reality is that, in a lot of situations, AI robotic systems can do certain tasks much better than humans (K. Hussain, 2018).

Medical Area

In medicine the utilization of AI is increasing, and its applications are getting more and more important. Fuzzy logic systems are used in diagnosing lung cancer, acute leukemia, and breast and pancreatic cancer, for example (Kamble & Shah, 2018). Machine learning techniques are used in radiology (Thrall et al., 2018). AI is used to alert doctors when a patient situation gets worse. In surgery AI can also be of help. It can also be used to prescribe medicines (Yeasmin, 2018). The list could go on, but what we can see is the perspective for the future is for AI to become integral part of medicine.

Transportation

In this area we can use road sensors as smart agents to detect accidents and predict traffic conditions. Artificial neural networks can be used for road planning and public transport planning. The autonomous vehicles that we are starting to see today are also leveraging AI technologies (Abduljabbar et al., 2019).

Weather forecasting

Nowadays climate conditions are predicted using AI techniques like neural systems. Based on past climate information that is fed into the neural systems, these machines are capable of examining the information and predict future conditions (K. Hussain, 2018).

COVID 19 Pandemic

AI techniques can be used in the Covid 19 pandemic for early detection and diagnosis of the infection, monitoring the treatment, contact tracing of the individuals, projection of mortality cases, development of drugs and vaccines, reducing the workload of healthcare workers, prevent the disease, and potentially for much more (Vaishya et al., 2020).

Smart Grids and Renewable Energies

AI techniques like expert systems, fuzzy logic and artificial neural networks have provided considerable evolution on power electronics systems and engineering. AI provides today powerful tools for design, simulation control, fault estimation and diagnosis, and fault tolerant control systems in modern power grid and renewable energy systems (Bose, 2017).

Customer Services

Most of the customer services today are already using AI systems, for example the chat bots on their website, or even some IVR solutions are already AI based (Kamble & Shah, 2018).

Entertainment

In our daily lives when we receive some music or playlists suggestions in Spotify, or when we receive the top 10 list of series or movies in Netflix, these are examples of AI systems applied to entertainment. There are AI algorithms that based on our preferences build those suggestions, and in the case of top lists, these are built on preferences from people near us or that have similar tastes (Verma, 2018).

IOT – Internet of Things

IOT networks are widely used nowadays, and like AI, they are present in a vast area of society. We can say that AI and IOT go hand in hand, as both are directly related and are creating a huge positive impact in our lives. Some applications of AI in IOT solutions are in home automation, oil and gas field production, smart booking for hotels (Mohamed, 2020). Most smart devices that make the IoT network are leveraging some type of AI technique.

Cybersecurity

The application of AI techniques in cybersecurity is the objective of this study, therefore in the following section we'll dive into the systematic literature review of AI and cybersecurity to acquire the knowledge about the state of the art of the usage of several AI techniques in the various cybersecurity dimensions.

3.2.7 Advantages and Disadvantages of Artificial Intelligence

Although AI is a very important technology now and for the foreseeable and not so foreseeable future, not everything about it is positive. Like anything, AI comes with a lot of advantages, but also some disadvantages.

Some of the advantages of using AI is that compared to a human being an AI agent has a much lower error rate. These systems are more precise, fast, and accurate than any human. Machines can work in a hostile environment as they're not subject to psychological, physical, or other human specific aspects that may affect performance. There are tasks that are known to be dangerous to a person, whilst a machine won't be affected by it, i.e., in case of an accident it doesn't involve any life loss. We can use as example space missions or oil and gas explorations. When it comes to execute repetitive and monotonous tasks a human can lose concentration and efficiency, whereas this is not something a machine will be impacted. Machines don't have emotions, so it may be that when it comes to rational decisions they can be more effective than a human (Pedamkar, 2020).

On the other hand, AI also comes with some disadvantages. Building AI machines, repairing or providing support requires very skilled professionals and it's very costly. The technology used by AI systems and mainly the storage to keep the huge amounts of data it processes is also very expensive. The machine learning processes allow the machine to know more and evolve, but not as much as a human being. Humans have a distinct characteristic which is creativity, which can't be found in a machine. One of the most discussed topics around the growth of AI has to do with impact it will have in employment of people, since a lot of jobs can now be performed by AI systems, leading to inevitable unemployment. The power of AI in the hands on unethical people can lead to its misuse or even to criminal activities (Pedamkar, 2020). Cybercrime is of course one of them, like we're studying in this work.

3.3 Artificial Intelligence & Cybersecurity - A Systematic Literature Review

One has used the PRISMA methodology to conduct the systematic literature review. PRISMA stands for: "*Preferred Reporting Items for Systematic Reviews and META-Analysis*" (Moher, 2009). It consists of a set of guidelines useful to conduct systematic literature reviews, critical literature analyses and meta-analyses (Moher, 2009).

3.3.1 PRISMA Methodology

The PRISMA methodology uses a set of methods to systematically search scientific papers and other scientific literature to conduct review-based studies. It is based on inclusion and exclusion criteria for any study being systematically assessed. According to the quality of chosen articles it either includes or excludes them from the study (Liberati et al., 2009).

Developed in 2005 by a group of 29 people, between scientists, authors and other specialists from several areas, after an extended consensus they agreed on a 27 item checklist and a four phase flow diagram, which resulted in a powerful template for researchers to use (Liberati et al., 2009).

The entire process is based on four phases that follow a simple workflow. The phases are:

1. Identification - Identify relevant articles, papers, and other scientific documentation based on a certain search strategy using the most common databases.
2. Screening – Use a very well-defined criteria to include relevant papers and exclude those that bring no value to the research.
3. Eligibility – Assess included articles for eligibility and exclude other for well justified reasons.
4. Included – Result list of papers that will be used as full source for the research

3.3.2 PRISMA Execution

The sections 3.1 and 3.2 of this text served the purpose of introducing the theoretical background for the two topics in study. From this theoretical background we can extract the keywords that were used in the search string one has built to find the scientific papers that are relevant for the study. The objective is to achieve a comprehensive understanding of the state of the art regarding the utilization of artificial intelligence in the cybersecurity area.
The research one has proposed to perform shall answer the following questions:

RQ1	What is the current status of research in this area?
RQ2	What are the major issues of AI in cybersecurity for businesses and governments?
RQ3	What kind of AI techniques are currently useful in this area?
RQ4	What are the advantages and disadvantages of applying AI techniques in this field?

Table 2 – Systematic Review's Research Questions

To answer these questions, and according to the PRISMA definition, one has selected the most relevant studies in this field. To conduct the search, a set of keywords that one has considered to be more relevant amongst the several concepts analysed in the theoretical background, have been chosen. One has opted to use only English words, and therefore the outcome of the search was mostly articles written in English. The ones written in other languages were excluded from the selection, according to the criteria defined in one's PRISMA execution. Keywords used are:

	Cybersecurity	Artificial Intelligence
Keywords	Cybersecurity	Artificial Intelligence
	Intrusion detection systems	Deep Learning
	Denial of Service	Machine learning
	Malware	Artificial Neural Networks
	Ransomware	Artificial immune systems

Table 3 – Systematic Review's Keywords

A specific search string was built to include the above words or terms with the objective of finding them in abstracts, titles or keywords of articles and other scientific papers. This choice of words assured that the results of the search would only retrieve data that's relevant for the topics in study. One was interested only in scientific documents that are recent, as this technology evolves very fast, and as such, only the most recent articles can ensure an up-to-date and relevant information. For this, one has set a filter to show only articles between 2018 and 2022, aiming to obtain accurate information about the current state of the art on the utilization of AI in the cybersecurity domain.
The search string used was: ("Artificial intelligence" OR "Machine Learning" OR "Deep Learning" OR "Artificial Neural Networks") AND ("cybersecurity" OR "Intrusion detection systems" OR "Malware" or "Ransomware") AND ("State of the art" or "Issues" or "Challenges" or "Techniques" OR " current status").
Note here, that besides looking for the keywords directly related to cybersecurity and AI, a Boolean query was also included to add the terms that appear in the research questions, with the objective to find articles that address the problems one is studying.
The search was conducted in November 2021 on the following scientific information resource databases:

Resource Database	Resource URL
Scopus	https://www.scopus.com/home.uri
Web of Science	https://www.webofknowledge.com/
Research Gate	https://www.researchgate.net/
Science Direct	https://www.sciencedirect.com/

Table 4 – Systematic Review's Resource Databases

Following the PRISMA methodology, the next step was to define the inclusion and exclusion criteria for the articles resulting from the mentioned search.

Inclusion Criteria	Exclusion Criteria
Any scientific article showing evidence of AI utilization in Cybersecurity	Papers focusing on Cybersecurity but without focusing on AI techniques utilization
Paper must be a peered reviewed conference or journal paper written in English	Articles not in English and duplicate papers
Paper is published in a scholarly journal with at least quartile two classification	Articles from quartile three or quartile four journals
Paper is published between 2018 and 2022	Articles published before 2018
	Non-academic or non-scientific papers (e.g., websites, magazines reports, newspapers, consulting articles, books, citations)
	Papers with titles outside the scope of this work

Table 5 – Systematic Review's inclusion and exclusion criteria

After inserting the search string in sources websites, as output one has got all the identified articles through database search, which resulted in a total of (n=7511) articles, i.e., we're in the identification phase of the PRISMA workflow. When moving to the screening phase, the first step is to remove duplicates. Here, (n=2115) articles have been removed, moving to the second step of the screening phase a total of (n=5396) records. In this second step of the screening phase, the inclusion and exclusion criteria has been applied: articles older than 2018, articles not in English, files not in pdf format, inaccessible articles, and the other mentioned exclusion criteria were applied. This resulted in the exclusion of (n=3452) articles, moving on to the eligibility phase a total of (n=1944) articles. In the first step of the eligibility phase, articles abstracts were further analyzed and the ones that didn't have direct relevance to the study were excluded. Articles focused on specific industry areas like IoT, or 5G networks, or power grids, disaster management or multimedia platforms for example, were considered very specific and non-relevant for this study, since one is looking for a wider applicability of AI in cybersecurity, and not studying its applicability in a specific industry, therefore (n=1597) articles have been removed. Thus, resulting in a total of (n=347) articles included in the qualitative phase, i.e., where one assesses the main text of the articles. In this phase one kept articles that have shown evidence of utilization of several methods of AI in several types of cybersecurity areas and addresses the topics of the research questions. The ones not contributing to answer the research questions have been excluded, resulting in a final list of articles included in the study of (n=46). This process is represented in the following workflow picture:

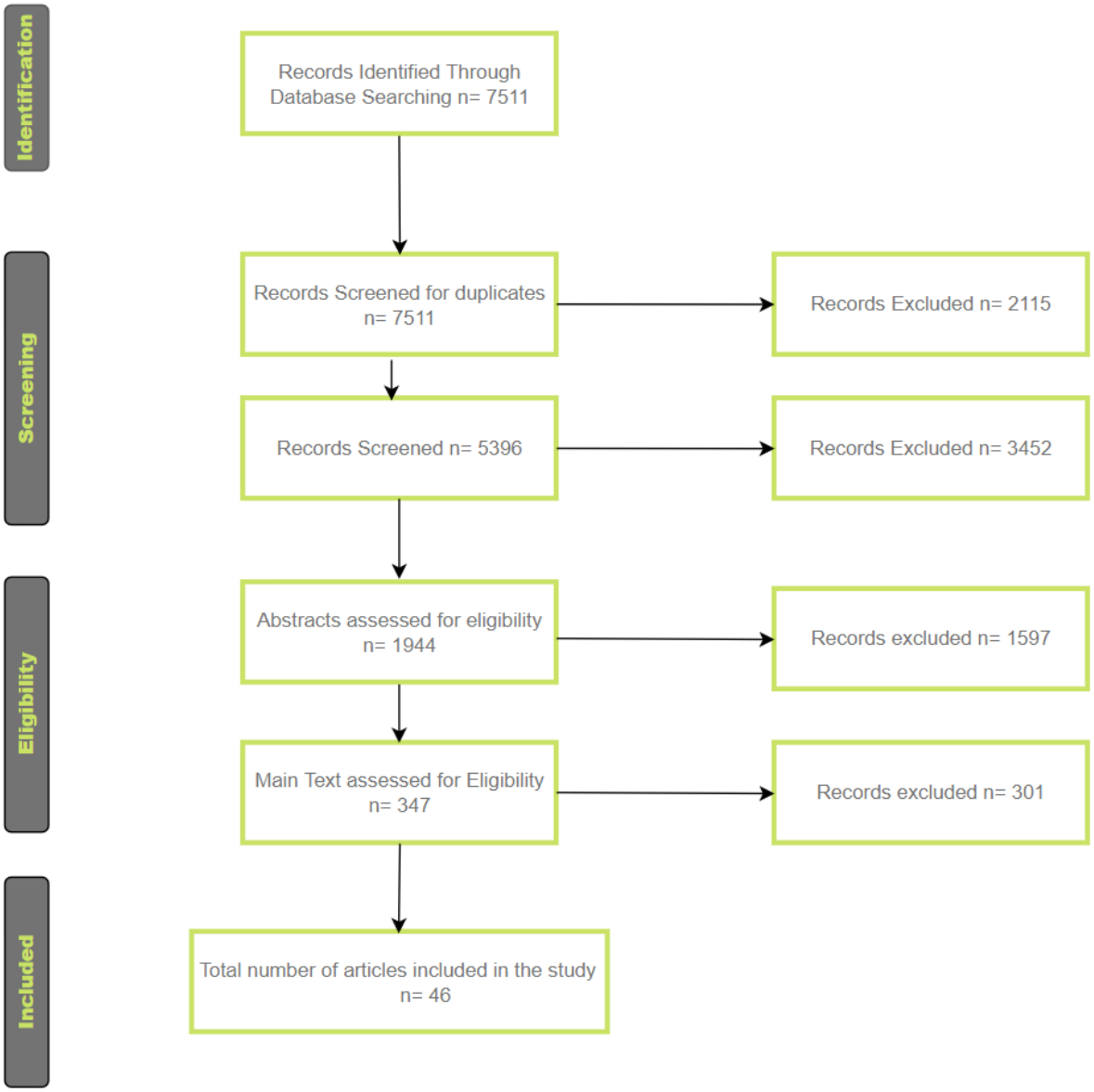

Figure 14 – PRISMA Execution

The output of this research, i.e., the articles moved to the included phase resulted in 40 journal articles and 6 conference papers, which are listed in the following table, along with a brief description of its contribution, conclusions, and gaps/future work suggestions.

#	Authors	Article	Contribution	Publication Type
[1]	(Alharbi et al., 2021)	Analysing the Impact of Cyber Security Related Attributes for Intrusion Detection Systems	Helps ML practitioners and cybersecurity specialists identify, select, and prioritize cybersecurity related attributes for IDS systems to build them more effectively.	Journal Article

#	Authors	Article	Contribution	Publication Type
[2]	(Ali et al., 2020)	A Systematic Review of Artificial Intelligence and Machine Learning Techniques for Cyber Security	Identifies the main ML classifiers used in the detection of cybersecurity attacks. Analyzes the state-of-the-art classifiers and presents a detailed taxonomy of the use of these classifiers in the detection of cyberattacks.	Journal Article
[3]	(Aljabri et al., 2021)	Intelligent Techniques for Detecting Network Attacks: Review and Research Directions	Provides an evaluation of contemporary network attack detection systems that use ML and DL models and presents some suggestions about future research directions in this field that can help filling some identified gaps.	Journal Article
[4]	(Alwaghid & Sarkar, 2020)	Exploring Malware Behavior of Webpages Using Machine Learning Technique: An Empirical Study	Analyzes machine learning methods that can be used to mitigate the behavior of malware in web pages. Provides recommendation and best practices to help in malware identification, using ML techniques like BART, RF, LR, SVM, ANN and CART to predict phishing attacks in emails.	Journal Article
[5]	(Arshi & Madhavi, 2020)	A Survey of DDOS Attacks Using Machine Learning Techniques	Analyzes and provides insights about DDoS attacks using ML techniques like Naïve Bayes, Multilayer Perceptron, and SVM, Decision trees. Provides suggestions for IDS systems based on Deep Learning that should be used to mitigate these types of attacks.	Conference Paper
[6]	(Aslan, 2020)	A Comprehensive Review on Malware Detection Approaches	Provides an analysis of malware detection approaches and current detection methods which use these approaches. Supporting researchers with a general overview of the malware detection approaches, pros and cons of each approach, and methods that are used, referencing the utilization of machine learning and deep learning.	Journal Article

#	Authors	Article	Contribution	Publication Type
[7]	(Atiku et al., 2021)	Survey On the Applications of Artificial Intelligence in Cyber Security	It provides a comprehensive review of the utilization of AI in the cybersecurity field. Presents benefits and challenges of AI application in cybersecurity. Concludes that with the current technology, AI systems are likely to improve protection in the cyberspace. Also states that the advantages outweigh the disadvantages.	Journal Article
[8]	(Barraclough et al., 2021)	Intelligent cyber-phishing detection for online	Presents a methodology combining blacklist-based, web content-based and heuristic based approaches, using ML algorithms which contributes to a more effective phishing attack detection.	Journal Article
[9]	(Basit et al., 2021)	A comprehensive survey of AI-enabled phishing attacks detection techniques	Contributes with a literature review of Artificial Intelligence techniques: Machine Learning, Deep Learning, Hybrid Learning, and Scenario-based techniques for phishing attack detection. Performs a comparison between different studies about the detection of phishing attack for each AI technique and examines the benefits and drawbacks of these methodologies. Also, provides a list of existing challenges related to phishing attacks and suggests research directions in this domain.	Journal Article
[10]	(Bécue et al., 2021)	Artificial intelligence, cyber-threats and Industry 4.0	Contributes with a discussion about the opportunities and threats of using AI technology in the manufacturing sector both from the defensive and offensive perspective. If analyses the utilization of machine learning and data mining techniques for intrusion detection systems presenting for each technique its strengths and weaknesses.	Journal Article
[11]	(Bello et al., 2021)	Detecting ransomware attacks using intelligent algorithms: recent development and next direction from deep learning and big data perspectives	Contributes with a survey about detection of ransomware attacks using intelligent machine learning algorithms. Presents some intelligent algorithm solutions from the big data perspective. Concludes that utilization of deep learning algorithms it's still at an early stage but are gaining a lot of traction in the detection of ransomware attacks.	Journal Article

#	Authors	Article	Contribution	Publication Type
[12]	(Cascavilla & Tamburri, 2021)	Cybercrime threat intelligence: A systematic multi-vocal literature review	Provides a study of the state-of-the-art techniques for cybercrime detection by using complex machine learning and deep learning investigation methods. It also provides insights in the utilization of these methods for cyber threat intelligence for deep and darknets. It contributes with a taxonomy of these methods mapping them to criminal activities, and risk indicators that are used for detection of these crimes.	Journal Article
[13]	(Caviglione et al., 2021)	Tight Arms Race: Overview of Current Malware Threats and Trends in Their Detection	Performs a review of existing studies related to malware and its detection techniques using AI. Presents a survey on malware analysis, evasion and detection, as well as ML applications on this field. Identifies several predominant techniques like evolutionary algorithms, shallow neural networks, reinforcement ML, DL and also bio inspired.	Journal Article
[14]	(Choi et al., 2020)	Using deep learning to solve computer security challenges: a survey	Provides a review of recent works related to the utilization of deep learning to tackle computer security problems. It addresses security issues like security-oriented program analysis, defending return-oriented programming attacks, achieving control-flow integrity, defending network attacks, malware classification, system-event-based anomaly detection, memory forensics, and fuzzing for software security. It concludes by analysis of these recent works that the utilization of DL techniques is still at an early stage of development.	Journal Article
[15]	(Feng, 2020)	Artificial Intelligence Cyber Security Strategy	Focus on strategy challenges related to AI utilization, suggestions for its use in cybersecurity and political trade-offs considerations. Challenges faced by organizations and governments with regards AI in cybersecurity are discussed in this article. It touches political and social challenges raising the attention to security controls using biometrics like facial recognition and its legal implications. Robotics utilization in forensics investigations and other applications is also discussed.	Conference Paper

#	Authors	Article	Contribution	Publication Type
[16]	(Gamage & Samarabandu, 2020)	Deep learning methods in network intrusion detection: A survey and an objective comparison	Provides a summary of recent research papers on the utilization of deep learning for IDS systems and introduces a taxonomy of deep learning models for IDS. Provides an evaluation and benchmark of four key deep learning models: feed-forward neural network, autoencoder, deep belief network and long short-term memory network, for the intrusion classification task on two legacy datasets (KDD 99, NSL-KDD) and two modern datasets (CIC-IDS2017, CIC-IDS2018). Identify some gaps in the research and suggest future research directions	Journal Article
[17]	(Gbenga et al., 2019)	Machine learning for email spam filtering: review, approaches, and open research problems	Presents a systematic review of the main spam filtering approaches based on machine learning. Analyzes the application of ML techniques in the SPAM filtering process for email service providers like google, yahoo and Microsoft. Provides a comparison of strengths and limitations of the several ML techniques applied to this field. Recommends the utilization of deep learning and deep adversarial learning as the future techniques that are more effective to handle this threat.	Journal Article
[18]	(Gibert et al., 2020)	The rise of machine learning for detection and classification of malware: Research, developments, trends, and challenges	Provides a description of ML techniques for malware detection and classification, focusing on deep learning techniques. Explores challenges and limitations of ML. Contributes to researchers with information in the malware detection field, current developments, and research directions. Highlights research issues, unsolved problems of the state-of-the-art methods.	Journal Article
[19]	(Hindy et al., 2020)	Utilizing Deep Learning Techniques for Effective Zero-Day Attack Detection	Proposes an intelligent IDS model, using deep learning, for zero-day-attacks detection. Model has a high detection accuracy overcoming known limitations of traditional IDS systems. Proposes future works that might focus on evaluating models using datasets that are specific for special purpose IDS systems like for IoT and critical infrastructures and adapting other ML techniques to be used in zero-day-attacks detection.	Journal Article

#	Authors	Article	Contribution	Publication Type
[20]	(Horan & Saiedian, 2021)	Cyber Crime Investigation: Landscape, Challenges, and Future Research Directions	A survey focusing on methods available for investigators on digital forensics and open-source intelligence to determine what kind of evidence each method provide, and which are more effective. Concludes that from all the methods available, natural language processing (NL), appears to be the most effective. Automation and ML are also critical in this field. Automation is helping investigators speed up their process of evidence collection and machine learning helps to identify and classify this evidence. Automation and ML are referenced as critical areas for future research to come up with more sophisticated solutions for cybercrime investigation.	Journal Article
[21]	(H. S. Hussain et al., 2021)	Artificial Intelligence in Cyber Security	Provides an overview on implementation of several AI technologies in various cybersecurity threats. Evaluates the prospect of enhancing defense mechanisms using AI. Analyzed several approaches and architectures, organizing them into the following categories: artificial neural, expert systems, smart agents, quest, computer education, data gathering, and constraint resolution. Emphasizes that AI technologies like Robotics, language understanding (NLP) and machine vision are very important and provide great capabilities for military applications, but not so much was found about its utilization in cybersecurity. Suggests that AI technologies related to management of data and information, specifically in the field of computer learning have the potential to significantly improve cybersecurity capabilities.	Journal Article

#	Authors	Article	Contribution	Publication Type
[22]	(Jamal et al., 2021)	A review on security analysis of cyber physical systems using Machine learning	Performs a security analysis of CPS using ML. Suggests a security framework to protect CPS from internal and external cyber-attacks. Identifies gaps and suggests directions for future research in the area. Identifies the need for integrated security for CPS and that several machine learning algorithms such as K-Nearest neighbor (KNN), Support Vector Machines (SVM) and Deep Neural Networks (DNN) are used to prevent DoS attacks, jamming attacks, time synchronization attacks, stealth time attacks, false data injection attacks in CPS. Concludes that there is need for more utilization of ML techniques in the design of CPS systems in an integrated approach for automating security.	Journal Article
[23]	(Kaloudi & Jingyue, 2020)	The AI-based cyber threat landscape: A survey	Analyzes existing studies on AI based cyberattacks. Analyzed studies are classified into five categories: next-generation malware, voice synthesis, password-based attacks, social bots, and adversarial training. It concludes that existing defense approaches are insufficient to tackle the increasing accuracy and speed of AI based cyberattacks. Propose a framework which provides insights into new threats, which does a classification of malicious utilization of AI in cyberattacks and provides a basis for their detection and prediction of future threats.	Journal Article

#	Authors	Article	Contribution	Publication Type
[24]	(Khan & Parkinson, 2018)	Review into State of the Art of Vulnerability Assessment using Artificial Intelligence	Due to the increase of AI techniques used in vulnerability assessments, the authors concluded that there was a need to perform research on the state of the art of this domain. Many AI techniques have been recently integrated in vulnerability assessment procedures, which resulted in the increase of quality, quantity and performance of the security threat identification and mitigation techniques. Techniques are found to be important in the aid of security experts by reducing their efforts, instead of fully replace them. Gaps have been found in the domain of automated knowledge acquisition and learning, limitations in computing power and memory, extracting human understandable results and autonomous resolution of identified issues.	Journal Article
[25]	(Lazic, 2019)	Benefit from AI in Cybersecurity	Addresses the role of AI utilization in cybersecurity and provides some recommendations on how organizations can benefit from it. Identifies benefits and drawbacks, but highlight those benefits outweigh the drawbacks.	Conference Paper
[26]	(Li, 2018)	Cybersecurity meets artificial intelligence: a survey	Analyzes recent studies related to combating cyberattacks using artificial intelligence, focusing on machine learning and deep learning methods. Analyzes and classifies the type of defense solutions that can be used. Identify several open topics that can be suggested as future research works like, AI based situational awareness for smart prediction and detection in the cyberspace, new and special AI algorithms for cyber security, specifically for big data intelligence, and also new security protection solutions for AI.	Journal Article

#	Authors	Article	Contribution	Publication Type
[27]	(Naik et al., 2021)	The impacts of artificial intelligence techniques in augmentation of cybersecurity: a comprehensive review	Focus on existing applications of AI techniques for analyzing, detecting and defend against cyberattacks. Identifies that AI is redefining cybersecurity and concludes that implementing AI techniques in detecting and preventing attacks is proving to have a huge potential. States that cost of detection and response of security breaches has decreased substantially when using AI techniques.	Journal Article
[28]	(Oseni et al., 2021)	Security and Privacy for Artificial Intelligence: Opportunities and Challenges	Focus on adversarial machine learning and propose a framework to demonstrate attack strategies against AI based applications. Analyzed several types of defenses that can potentially protect the AI applications against these attacks. Limitations have been identified namely in the fact that deep neural networks are subject to transferability of adversarial attacks. Researching this phenomenon becomes therefore critical to develop more robust ML models.	Journal Article
[29]	(Pachhala et al., 2021)	A Comprehensive Survey on Identification of Malware Types and Malware Classification Using Machine Learning Techniques	Provides a description of techniques used in ML for malware detection and classification. Highlights challenges and limitation of such ML techniques. Investigate research trends and progress in this field, focusing on deep learning approaches. Identify research problems related to state-of-the-art methods. Provides suggestions in future research direction on malware detection using AI.	Conference Paper
[30]	(Rosenberg et al., 2021)	Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain	Provides a summary of recent research regarding adversarial attacks on security solutions based on ML technologies, highlighting its risks. Applications of attack and defense methods using AI are categorized. Identifies and discuss the challenges of implementing end to end attacks on security solutions that use AI and map them into a taxonomy that helps the community in future research directions. Identifies gaps in metrics to measure robustness of classifiers for several attacks, and defense methods that are robust against unknown adversarial attacks.	Journal Article

#	Authors	Article	Contribution	Publication Type
[31]	(Sagar, 2019)	Providing Cyber Security using Artificial Intelligence – A survey	Highlights the need for the development of cybersecurity skills and how AI can be used to improve the cybersecurity landscape by using AAN and ML algorithms. Concludes that is necessary to maintain the technologies up to date according to latest requirements because AI is evolving fast.	Conference Paper
[32]	(Shaukat, Luo, Varadharajan, & Hameed, 2020)	Performance Comparison and Current Challenges of Using Machine Learning Techniques in Cybersecurity	Contributes with a detailed review of machine learning techniques in cybersecurity. Reviews different ML techniques with the objective of understanding the developments made in defense mechanisms for potential security threats. Concludes that there is the need to develop robust machine learning models to tackle adversarial approaches. The emphasis must be on training ML models in adversarial settings so these robust models can be developed.	Journal Article
[33]	(Shaukat, Luo, Varadharajan, & Member, 2020)	A Survey on Machine Learning Techniques for Cyber Security in the Last Decade	Focus on the challenges presented to ML techniques when used to protect the cyberspace, by conducting a literature review on ML techniques used for intrusion detection, spam and malware detection on computer networks and mobile networks on the last decade. Concludes that is impossible to provide one recommendation for all the attacks using a single ML model. Several criteria like detection rate, time complexity, time to detect zero-day attacks and accuracy of an ML model must be considered when opting for a certain model do detect and prevent a cyberattack. Identifies a gap in literature where there is no study on ML techniques on both mobile and computer networks in the same article.	Journal Article

#	Authors	Article	Contribution	Publication Type
[34]	(Truong et al., 2020)	Artificial Intelligence in the Cyber Domain: Offense and Defense	Provides an overview on how AI is used in cybersecurity from the perspectives of both offense and defense. Discusses the impact of AI in the cyber domain, provides a survey of AI applications covering a wide range of attacks, discusses security threats related to adversarial AI. Concludes that that DL is a trend in the detection of malware analysis and classification, SPAM filtering and phishing attacks. Highlights that the combination bio inspired AI techniques together with ML/DL are presenting very promising results and should continue to be a trend in research. Stresses the challenge of the adversarial attacks on AI models which are leading to the appearance of autonomous intelligent malware.	Journal Article
[35]	(Veiga, 2018)	Applications of Artificial Intelligence (AI) to Network Security	Provides an overview about the needs for evolution in cybersecurity methods and how AI techniques can help. Presents a the state-of-the-art on AI network security techniques and analyses what can be the foreseeable future of applying AI in network security. It argues that currently only ML as a branch of AI is being successfully applied to solve a small part of this problem. Suggests that supervised ML has delivered some good solutions, but the research on unsupervised ML models should be the goal so it reduces human interaction as much as possible.	Journal Article
[36]	(Wiafe et al., 2020)	Artificial Intelligence for Cybersecurity: A Systematic Mapping of Literature	Author argues that there is a lack of literature on AI methods do combat cybercrime, hence it provides a systematic review of 131 papers utilizing quantitative and qualitative approaches. Concludes that AI techniques have contributed to fight cybercrime showing a considerable improvement in IDS systems. Observed that there is a reduction in computational complexity, ML models training times and false alarms. Most studies are focused on IDS systems and the method most used is support vector machines. Suggests that research should focus on other techniques to enhance the research in this field.	Journal Article

#	Authors	Article	Contribution	Publication Type
[37]	(Yamin et al., 2021)	Weaponized AI for cyber attacks	Contributes with a study on recent cyberattacks that use AI base techniques and identifies several mitigation strategies. Identified the main techniques being used for weaponized AI systems and what are the probable future scenarios that must be created to tackle and control such attacks. Concludes that to control the advancement on the weaponization of AI, there is a high political aspect and that superpowers must find compromise and reach to an international consensus. There must exist a manageable research and development of controlled AI cyber arms.	Journal Article
[38]	(Zeadally, 2020)	Harnessing Artificial Intelligence Capabilities to Improve Cybersecurity	Studies the potential of AI technologies to improve cybersecurity solutions identifying its strengths and pitfalls. Provides suggestions for future research opportunities based on identified gaps. States that the most recent AI security solutions are mainly focused on ML intelligent agents that are designed to distinguish between attack traffic and legitimate traffic. Highlights the likelihood of cybersecurity solutions evolving from intelligent agents acting humanly to thinking humanly. Identified gap is the need to develop intelligent algorithms to act on this field.	Journal Article
[39]	(Zhao et al., 2002)	A Review of Computer Vision Methods in Network Security	Contributes with a review of computer vision methods on phishing detection, malware detection and traffic anomaly detection. Concludes that some commercial solutions providers are evaluating the feasibility of combing machine visions methods with traditional ML models. For detecting zero-day-attacks machine vision methods have shown some advantages, as well as on accurate phishing detection systems. Highlighted gaps are that to improve research in this area it's necessary to establish larger and more recent datasets to define benchmarks and be able to compare the different solutions. Research direction should be on the development of hybrid solutions combining these methods that are more resilient to adversarial attacks.	Journal Article

#	Authors	Article	Contribution	Publication Type
[40]	(Dutt, 2020)	Immune System Based Intrusion Detection System (IS-IDS): A Proposed Model	Proposes an intrusion detection model that mimics the natural immune system considering both its layers, innate immune system, and adaptive immune system. It proposes a statistical modeling-based anomaly detection as the first layer of an IDS. Conclusion is that these model yields 96% true positive rate of detection of real time traffic attacks.	Journal Article
[41]	(Farzadnia et al., 2021)	A novel sophisticated hybrid method for intrusion detection using the artificial immune system	Author states that artificial immune systems as a good prototype for developing machine learning, is a good option to design IDS systems. The negative selection and the danger theory, inspired by immunity responses of the human immune system are being widely used in this area, therefore the study proposes an hybrid model that includes two defensive lines for attacks, by using these two mechanisms of AIS. The study concludes that they can improve the IDS capabilities of state-of-the-art IDS by using this approach.	Journal Article
[42]	(Louati & Barika, 2020)	A deep learning - based multi - agent system for intrusion detection	Contributes with a model for a DL based multi-agent system for IDS combining the optimal features of a multi agent system approach with the precision of DL algorithms. The model is then compared to other more conventional ML approaches for IDS and other multi agent systems, showing that this hybrid approach for IDS achieves higher and faster detection rates when compared to more conventional IDS. Suggest future work on applying this model to real-time network traffic, and extending it to cloud computing, fog computing and IoT.	Journal Article

#	Authors	Article	Contribution	Publication Type
[43]	(Nguyen & Reddi, 2021)	Deep Reinforcement Learning for Cyber Security	Presents a survey on DRL approaches specifically developed for cybersecurity. Addresses DRL based security methods for CPS, autonomous intrusion detection techniques, and multi-agent DRL-based game theory simulations for defense against cyber-attacks. Concludes that utilization of DRL techniques is an emerging area in terms of security solution for cyber-physical systems (CPS). More models and other simulations for this type of systems are suggested as future work. In addition, its found that for IDS systems there is a lot of work with RL methods, but not many using DRL, hence the suggestion on future work in this are as well.	Journal Article
[44]	(Shenfield et al., 2018)	Intelligent intrusion detection systems using artificial neural networks	Introduces a new architecture approach to detect malicious network traffic using ANNs specifically for utilization in deep packet inspection-based IDS systems. This architecture assures 98% of accuracy on average, proving this solution to be robust, accurate and precise. It has the potential to improve the utility of IDS systems not only on network traffic, but also on CPS and smart grids. Future work can be done on solutions to detect cross-site scripting attacks and SQL injection attacks on web applications	Journal Article
[45]	(Suliman et al., 2018)	Network Intrusion Detection System Using Artificial Immune System	Proposes a model that uses AIS for detection of computer networks intrusions. Uses classification methods for each feature on a computer network, like protocol used, type of connection, type of service, which is then used by the AIS model to distinguish between valid connections and attack connections. Results show this model to have a high success rate in identifying attack connections. Author states that results are as good as other models found in literature.	Conference Paper

#	Authors	Article	Contribution	Publication Type
[46]	(Zhang et al., 2019)	Intrusion Detection for IoT Based on Improved Genetic Algorithm and Deep Belief Network	Presents an IDS model based on an improved genetic algorithm (GA) and deep belief network (DBN). Proves that DBN can process efficiently high complex data and all the results are good. The method has the advantage of having a high classification accuracy which can result in detection rates above 99%.	Journal Article

Table 6 – PRISMA results table – included articles

3.3.3 PRISMA Results Analysis

Having completed the research of the required information for the study, following the PRISMA methodology which has been thoroughly described in the previous section, now one performs the analysis of the results of this research, i.e., one must analyse each of the included articles with the purpose of retrieving the main contribution of each work and find the answers for the research questions.

AI techniques currently useful in this area

From the literature analysed one can conclude that there is a lot of research regarding the application of artificial intelligence techniques in the domain of cybersecurity, being machine learning the field most used in the state-of-the-art security applications [10].
Literature also suggests that there are certain techniques and technologies that are more used, and that not all AI techniques are being applied. In what regards the most used techniques, one finds that most literature tend to focus considerably in applying AI in intrusion detection systems (IDS) [10], especially using deep learning techniques (DL), like can be seen in [14, 16, 36] and [42]. In [40, 41] and [45] there are also models for IDS solutions but using artificial immune systems (AIS). Artificial neural networks (ANN) [44], and deep belief networks (DBN) [46] models for IDS, are also used.
Deep learning has shown to be a critical technique that has impact and good results in several domains in cybersecurity. A survey performed in [14], covers several types of security problems that can be successfully solved using deep learning: security-oriented program analysis, defence against return-oriented programming attacks, control flow integrity achievement, protection to network attacks, classification of malware, anomaly detection based in system event logs, memory forensics and fuzzing software for security solutions.
Deep learning has also had recently a high impact in solving security problems in some type of industries, smart grids and IoT being good examples [22].
Several other security threats that have been discussed in the theoretical background are also addressed by different type of AI techniques. In [27] we can find information stating that fuzzy expert systems have been recently a hot topic in research and being applied in some use cases of several types of network attacks. Comparisons between these fuzzy expert systems and others like ANN,

fuzzy neural networks, genetic algorithms, are also being researched, having in mind threats like viruses and worms' type of malware.

Deep learning algorithms are also used to detect ransomware [14]. Machine learning classifiers like support vector machine (SVM), random forest (RF) and decision trees (DT) are frequently used in the detection of malware cyberattacks [2, 3].

Another well-known and widely used ML classifier, a supervised learning model, is Naïve Baynes, which is being used for DDoS attacks, zero-day attacks, phishing attacks, and botnet attacks [3].

[5] adds that support vector machine and decision trees are also used for DDoS attacks.

In [6] there are discussions about malware detection, namely new types of malwares that cannot be detected using old signature detection schemas. The dynamic features like obfuscation and polymorphic techniques of the novel types of malwares, can also be combated with machine learning and deep learning algorithms.

One can see the malware challenge being generically addressed in [7], and in [13] is highlighted that the most used techniques are evolutionary algorithms, shallow neural networks, reinforcement machine learning, deep learning, bio-inspired computation, and swarm intelligence. Here is also concluded that deep learning methods for malware detection have shown better performance than other techniques like random forest and Naïve Bayes.

Phishing is another important threat we face today, and AI is also used for preventing and detect phishing attacks [8]. According to [9], machine learning, deep learning, hybrid learning, and scenario-based techniques are useful techniques in phishing attack detection. Moreover, one has found in [39] that computer vision methods can also be used in phishing detection. Computer vision is an AI technique for which very few references were found, still in [39] there is also have evidence of its use for malware detection and traffic anomaly detection.

Email spam, another cybersecurity problem, can also be fought with AI. Detection and filtering of spam emails can be optimized using machine learning, and [17] suggests that deep learning and deep adversarial learning are the techniques that can better tackle this menace.

So far, one has been addressing the cybersecurity challenges from the threats and its defences perspective, but additionally it can be found in literature that AI is also being integrated into cybersecurity functions, like biometric based login systems, conditional authentication and access techniques, detecting threats and malicious activities with predictive analytics, using natural language processing to improve learning and analysis of some common security functions [25]. CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) a very well-known function used in web sites authentication, is another example of an AI mechanism used in cybersecurity [27].

Vulnerability assessments is also a domain where AI is being used, using machine learning algorithms as well as expert systems that use fuzzy logic techniques [24].

AI forensics is also an emerging area where natural language processing is showing to have several applications [20]. Still in the field of cybercrime investigations, according to [20], automation and machine learning are showing considerable advances.

Generically speaking, across all literature, but emphasized in [32] and [33], it's seen that traversal to most of cybersecurity threats, the most dominant use of machine learning techniques classifiers are support vector machines, Naïve Bayes, random forest, artificial neural networks, decision trees, for supervised learning and K-Mean, Deep Belief Networks for unsupervised learning.

Status of the research in this area

When evaluating the state-of-the-art research in this area, one is more concerned in what regards security models and ontologies that can be found in literature and how the industry and researchers are evaluating this thematic, and not so much on the specific AI techniques and its applications to a specific security problem.

Some researchers have raised some concerns towards ML application in identifying zero-day-attacks, specifically showing some scepticism about its effectiveness in real attacks situations [10]. Argument being that ML systems are trained using historical data and therefore it may not be effective on operations that face new attacks or certain data pattern unobserved yet [10].

We've seen previously how deep learning algorithms are getting so much relevance within the research community, being used to tackle several types of threats. However, a survey conducted in [11] concludes that when it comes to detect ransomware in big data architectures, the application of deep learning seems to be a virgin area where a lot of challenges are being raised. Big data architectures are a preferred target for ransomware attacks due to the potential it has in terms of data value and the amount of money ask that can be put into a ransom. This fact leads to the obvious conclusion that being this an area where the application of intelligent algorithms is at a very early stage, it's critical to invest a lot more effort into developing better security models [11].

In terms of cybercrime detection, using complex ML and DL approaches, it can be found in [12] information highlighting the application of engineering activities on surface, deep and dark webs, presenting the state-of-the-art of research in this specific aspect.

Despite all the good metrics and results observed in using ML to detect threats, the survey in [13] states that the relation between deep learning and cyber security related activities is far from being mature and considers that quantifying the real capabilities and effectiveness of using machine learning methods is still considered an open point. It continues by arguing that the fact that ML based approaches require manual labelling, which is a time-consuming task, prone to errors and somehow costly, puts a limitation tag on the whole concept.

This same work, [13], mentions that there are many surveys focused on ML applications for specific aspects of malware, like detection or evasion mechanisms, and some other studies are covering malware that is targeted to specific domains like Internet of Things (IoT) or mobile networks, smart grids, and other types of infrastructures. It goes on by showing in their conclusions that in some scenarios where the attackers themselves use ML algorithms to evade malware detection systems, in some cases, these techniques are not fully effective against adversarial attacks.

Still on the malware topic, in [18] it's shown a conclusion that the state-of-the-art research on the utilization of the different type of machine learning methods for its detection is gathering a lot of focus and attention due to identified limitations on keeping up with the fast evolution of novel malware types.

Another relevant finding regarding the status of research in this domain is the fact that, according to [23], there are a lot of studies published focusing on the advancements of AI, but not so much attention has been provided to the dangers and challenges it brings. The utilization of AI, as one will present later in this text, is not full of virtues and it comes with some disadvantages as well. In [23] it's stated that there is limited research about the ways that AI can be used as a malicious tool by attackers. It adds that, regarding the possible utilization of AI technologies for malicious purposes, there is a lack of systematic knowledge. This was the conclusion of a study performed by ESET, a

Slovak internet security company, targeted to IT managers in some of the most relevant companies in the US, UK, and Germany, about their main concerns regarding AI in the cybersecurity domain.
The authors in [23] also contribute with some information about relevant works done recently which contribute to the current landscape of security models using AI. As an example, in [23] it can be found a model where AI is used to build proactive and reactive mechanisms against malicious cyber activities. It also mentions a model which uses computational intelligence approaches in building intrusion detection systems. A model based on deep reinforcement learning to protect critical physical systems is also referenced in [23].
It has also been found in literature, [26], that it's important to build some artificial intelligence models, namely safe distributed ML/DL intelligent systems, and find ways of doing machine learning classification over encrypted data.
To conclude the brief analysis regarding the status of existing research on AI and cybersecurity, it is relevant to mention how the security industry is leveraging this topic, and in [25] we can find that the most businesses, or at least the ones showing more concerns about the cybersecurity thematic, are already integrating AI based cybersecurity tools and solutions from some of the most important players in the security market like Check Point, CrowdStrike, FireEye, Fortinet, LogRhythm, Palo Alto Networks, Sophos and Symantec, just to name a few.

Advantages and disadvantages of applying AI in this field

In most of the subjects in any area of life and technology we can find advantages and disadvantages. When it comes to the utilization of AI technologies in cybersecurity this is no different.
In [7, 21, 25, 33] and [34] one has found an extensive review of advantages and disadvantages of AI in the domain of cybersecurity.
Starting with the advantages, in [7] it’s stated that organizations that have implemented AI in their cybersecurity strategy have seen significant benefits and evident return of investment on their cybersecurity tools.
In Siemens AG, the AI based tools were able to estimate 60.000 potential assaults per unit time. As a result of this, the management of the system was possible to be achieved with a team of under just twelve members, without any decrease in performance.
Also, in [7], its concluded that about 64% of company administrators revealed that the adoption of artificial intelligence cut down the costs of identifying and reacting to breaches. It became evident that some limitations of earlier security technologies are resolved by this new smarter technology [7].
Most of companies and organizations also say that since they started using AI, their cyber security professionals became more efficient and more accurate in performing their roles [21]. AI contributes to a reinforcement of company’s digital management strategies used to fight cybercrime, consequently helping them keep their businesses and their customers more secure [21].
According to [25], by using AI, companies state that the time necessary to detect attacks and breaches had a 12% reduction and remediating the breaches or deploying patches in response of these attacks also takes 12% less time. Using AI in their cybersecurity strategy, organizations can reapply prior threats patterns to identify new threats and deviations, this resulting in less time and effort previously spent in incident investigation and threat remediation [7, 25].

For example, AI can learn about normal user behaviours and with time will understand that if a user logs in in a system in one country and after a little while the same user is attempting to login in a different country it's an unusual pattern which then can trigger an alert [25].
The gathering of new information allows AI systems to improve its own functionalities and strategies. This can be seen as a kind of predictive behaviour which empowers the security teams with an advantage that becomes crucial to stop attacks before they happen instead of having to remediate them afterwards [25]. In essence AI can detect new and very sophisticated changes in attacks flexibility and can learn over time how to better detect and respond to threats [34].
The scale at which the machine learning techniques can analyse and filter massive amounts of data, like systems logins, event logs, network traffic patterns, and any type of computer usage information, is in no way comparable to what any human can do [34, 25]. For any person to perform such a task would be impossible, whereas AI systems do this quickly, without any effort around the clock, i.e., 24 hours a day, 7 days a week all year round [25]. This also means that the AI protection systems, while being active 24/7, and being capable of processing so much information, can provide not only a continuous protection, but also a response to cyber threats which would require many days or months or in some cases even years for a human being to recognize [21].
This huge capability of processing large amounts of data and, very quickly going through structured and unstructured data, comprehensively reading statistics, words, and phrases, also means that organizations, either private corporations or governments, can save a lot of time and money. From the governments point of view this mostly means saving tax money as well as national secrets [25].
Nowadays hackers are constantly trying to find solutions to trick the machines, breaching through security systems using mechanisms still unknown. This means that it can pass months before an organization realizes it has been compromised, and by then the hacker is probably already gone taking with him the stolen data. If AI is used, the systems can be actively collecting lots of data just waiting for a hacker to perform his actions. AI can detect abnormal behaviours that hackers commonly present, therefore preventing a considerable loss [25].
AI contributes to cybersecurity by changing the paradigm from a manual reaction/remediation logic to an automated prevention/remediation logic [25].

There are also some challenges and disadvantages that come with the utilization of AI.
To build an AI system it's required a large number of data input samples, and it´s very time consuming to gather and process all these samples [7, 34].
Systems that use AI are very resource intensive, i.e., they require powerful hardware with high processing power, memory, and storage to process such large data sets [7, 21, 34]. This makes it very costly, meaning not all organizations will have the means to implement it [7, 21].
AI systems still require human oversight because they cannot function in a fully autonomous way, as technology tend to detect too many false positives and on the other hand it also let several attacks go undetected. So, although in many aspects it does jobs that a human cannot do, still it cannot fully replace a human being, especially when it comes to decision making activities that objectively protect the organizations from some types of cyberattacks [25].
The people needed to operate AI systems are highly skilful resources, which sometimes are hard to find and are also expensive [7]. So, effectively AI requires more technology and financial resources than traditional security solutions non-AI based [25].
Another disadvantage of AI is that these technologies are also available to the attackers, which provide the criminals with tools that make them more dangerous and effective [21]. AI can serve as a

new weapon in these cybercriminals' arsenal empowering them to trigger more sophisticated cyberattacks [7, 25]. Some of these new attackers' tools include methods like adversarial inputs, model theft and data poisoning, which are aimed to target the AI defence system itself [7, 33, 34]. Adversarial AI is defined as "*something that causes machine learning models to misinterpret inputs into the system and behave in ways that's favourable to the attacker*" [25].

False alarms generated by AI systems is also a challenge that often disrupts businesses because it can cause delays in responses that are critical, i.e., dealing with false alarms divert the attention of resources which could be applied to real situations [7, 34].

According to [7] and [25] there is one important disadvantage of AI that is often neglected, which is related to the human beings, specifically the challenge is that in the companies where AI is adopted, employees may become less conscious about preventive behaviours, lowering their guard is some occasions, and this element of complacency can result in some risks.

To conclude, one final aspect is raised in [33], which is related to the ML models and the data sets used to train them. According to this study, most datasets currently used are outdated and they don't seem to have the amount of data that would be necessary to perform at their best.

Moreover, since cybercriminals are evolving their methods very fast, we must ensure the most efficient models for detection, and besides the use of the right dataset, [33] argues that there should be an ML model specifically designed to deal with a specific type of cyberattack, instead of relying on generic models for detection of different types of attacks. According to this study this is currently seen as a limitation of AI utilization.

Major issues related to AI in cybersecurity for businesses and governments

When looking at the main issues that AI utilization presents to organizations and governments, we may consider that some of the disadvantages seen previously also fit in this section, however, in literature one has found some aspects that have more impact to the organizations itself, therefore they are addressed here as separate topic.

One of the most evident issues organizations face is related to costs, that is, to address cybersecurity issues, businesses and governments have been spending over the years huge amounts of money in traditional security solutions, and with the rise of AI those costs became even higher, because they now need to invest in AI based solutions, which, like seen before, are very costly and take more time to implement [1].

Banks have suffered losses in the order of trillions of dollars in frauds that are mostly caused by attacks using AI, and this leads to the need of spending more money in their defence systems to implement AI based protections. Here the payment card industry has been paramount in helping the banks [21]. In this problematic, the increase in AI based attacks like malware [6], phishing and spam emails scams [8, 9] have the most impact, resulting in huge financial losses for organizations.

It's been observed that it's difficult for the companies to design AI intelligent systems that does not present any negative effect when performing their cybersecurity tasks [27].

The lack of scalability of AI based security systems is also a problem for organizations, and with the evolution of these technologies and how they can do work that used to be done by humans, it brings the challenges for organizations on how to handle the human problem of having to let go people [27].

If in one side some people are replaced and need to leave, on the opposite side, adopting AI also means that the security analysts that are still required need to be trained on the new technologies,

and potentially some must be recruited, which for some companies can also be a challenge, and of course also brings additional cost [25].

Besides this human-resources problem, AI is also bringing problems related to ethical and legal aspects [27]. As an example, the use of facial recognition systems if done by governments can be considered illegal or a violation of human rights [15].

Another type of concern is related to the use of automation in the field of digital forensics which also brings some issues regarding legal assumptions and its implications [20].

Other social and political challenges arise with the use of AI. In [15] is mentioned that during the COVID-19 pandemic crisis, the UK and other countries decided to use AI based applications related with the testing and tracking of people so they could control the spreading of infections. Although these may not be strictly related to cybersecurity in the technological sense, it's still a challenge brought by AI applied to an important security issue.

The utilization of AI biometrics authentication brings a new type of challenge which is the fact that these systems can keep records of biometric data for a very long time, it can even last for a lifetime. This raises the need to create the right laws to protect data, make sure it isn't released without consent, putting an extra concern in governments to think very carefully how to treat biometrics data, making its security a priority in their security strategies [15].

By adopting AI, organizations also need to develop new governance models so AI systems can deliver long term improvement when comparing to old systems. Implementing a good governance model is critical to ensure that AI systems deliver the expected outcome and are not compromised [25].

Besides economic, social, governance and political issues, companies also face technical issues related to the implementation of AI. In [15] it's stated that it's impossible to develop any software or other AI solution that is flawless. Some errors caused by AI systems can have consequences as bad as the loss of a human life, which surely influence governments strategy in using AI [15].

More technical challenges have to do with the difficulty in using the right dataset and the right ML or DL algorithms to identify the different type of cyberattacks. This has shown to be a difficult issue for organizations to solve [3]. Until now, there is no known algorithm that can identify and classify every attack with high accuracy [3]. It's recognized by experts that it's not reasonable to adopt just one algorithm as a universal model, which makes this topic another important challenge faced by organizations [3]. This situation results in the inability to develop more robust algorithms because there is limited availability of datasets that allows the detection of different network attacks [3].

To conclude one must bring up the issues that AI brings to cyber physical systems. Organizations that manage CPSs find in AI some new challenges which can threaten even more their security. The use of AI by attackers are a major concern for these organizations because these cyber security systems often have an interconnected nature meaning that a single vulnerability can have a domino effect in the entire facility [21, 22, 23]. The AI type of attacks on CPSs have not been thoroughly studied but some advanced threats can compromise these facilities to the point of taking over their control systems, insert botnets that can attack water distribution systems, taking down gas distribution facilities and a lot more [23].

4 DEVELOPMENT OF A SECURITY FRAMEWORK USING AI TECHNOLOGIES

Following the literature review analysis which resulted in finding the answers for the research questions, one is now in conditions to build the framework proposal.

4.1 KEY FINDINGS AND ASSUMPTIONS

One starts by briefly summarize the main findings that were thouroghly described in the previous section using the appropriate references, before going to the framework proposal itself. For this purpose, these findings are consolidated in a more graphical way, which aims to serve as the base assumptions for the framework development that follows. Figure 15 is therefore showing what AI branches are currently being used in cybersecurity, within the several branches it focus on the most used branch which is machine learning, by highlighting its techniques as well as the most used ML classifiers. Moreover it shows the most common security threats and cybersecurity domains that are being addressed with AI, together with the most common industries where the thematic is applicable. Subsequently it presents the technical and business implications of AI utilization in the cybersecurity landscape.

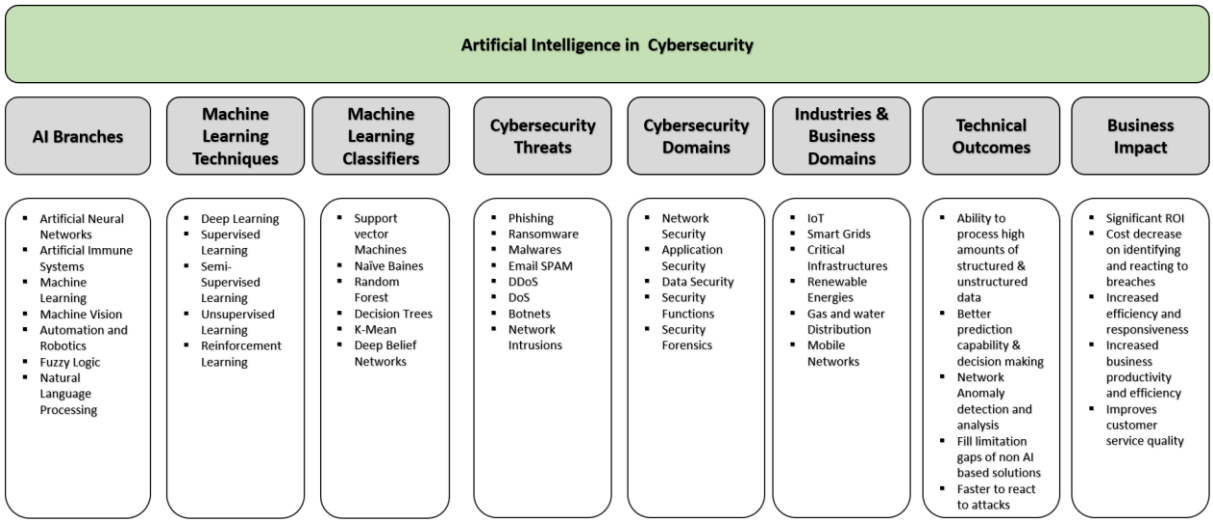

Figure 15 – Artificial Intelligence in Cybersecurity

Next, one presents a graphical representation that maps the correlation and applicability of the several AI techniques in cybersecurity. In the below figure can be seen that for some areas there is a direct applicability between an AI technique and a cybersecurity domain or threat, which is represented with a green tick. The empty cells mean there was no reference found in literature about its direct applicability, so it's areas that potentially can be considered for further investigation. It's also represented a few applications that can be considered a grey area as per the literature findings, that is, cases where the effectiveness of a technique is challenged, like ML for Zero Day Attacks. For the specific case of big data security, the question icon means that the usage of that specific technique is still considered a virgin an area, therefore requiring further investigation efforts.

Figure 16 – Mapping of AI techniques with cybersecurity domains and threats

4.2 FRAMEWORK PROPOSAL

To propose a security framework using AI technologies, one cannot ignore that, like we've seen in the theoretical background study, the industry already has several security frameworks which serve as reference for organizations to develop solid and compliant security solutions. Based on this, proposing a new framework inevitably will be built on top of existing concepts and foundations already used in the industry, but adapted to the purpose one wishes to fulfil; A model that uses AI technologies but at the same time being realistic and in line with the known best practices. The prosed framework is presented in figure 17, as a high-level view, which comprises of five main steps: Strategy, Design, Implementation, Operations and Roadmap:

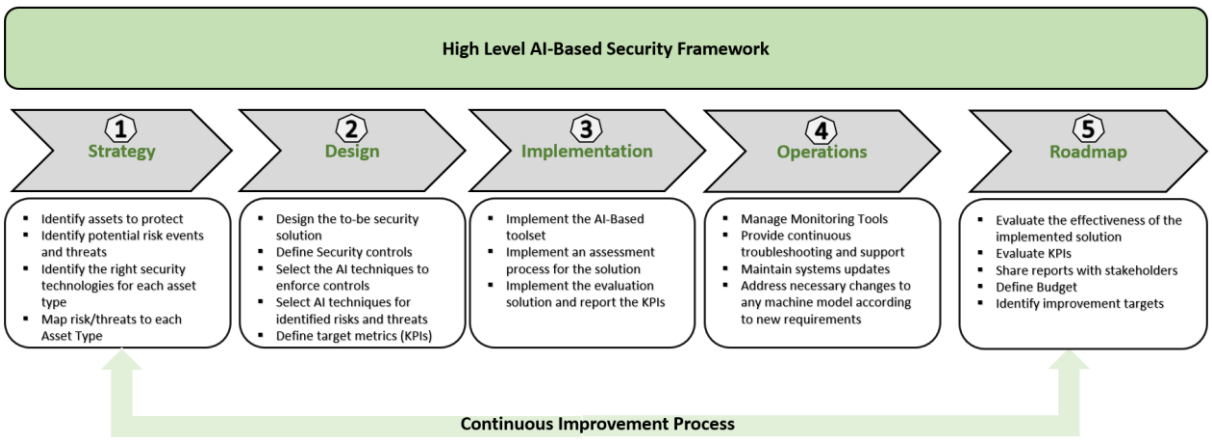

Figure 17 – High Level AI-Based Security Framework

In the next sections a drill down of each of the five steps is presented.

4.2.1 Strategy

Proposal for the strategy phase is comprised of four essential steps. Organizations shall start by identifying the assets that need to be protected, then identify potential risk events they may be subject to, and the consequent threats. Having defined this, organizations can identify and list the right security technologies for each of the asset type. This can be represented in a model like the one in figure 18.

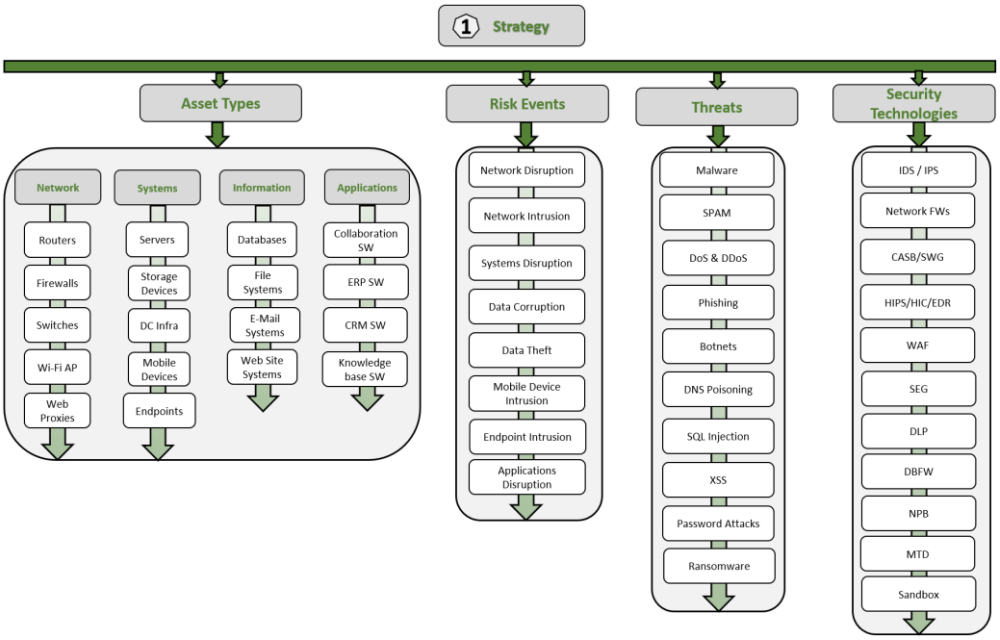

Figure 18 – Proposed Model for the Strategy Phase

Having this model present, organizations can do the exercise of mapping the existing threats to each asset type. One's proposal is presented in the following table:

Mapping Matrix

Asset Type	Asset	Is Subject to Risk Events	Is Subject to Threats	Applicable Security Technologies	References
Network	Routers Firewalls Switches Wi-Fi AP Web Proxies	Network Disruption Network Intrusion	DoS/DDoS DNS Poisoning Malware Botnets Password Attacks	IDS/IDPS Network Firewalls NPB Sandbox	(Yadav, 2020) (Bosworth et al., 2014) (Gartner, 2021b)
Systems	Servers Storage Devices DC Infra Mobile Devices Endpoints	Systems Disruption Mobile Device Intrusion Endpoint Intrusion	DNS Poisoning Password attacks Malware Botnets DoS/DDoS Ransomware	Network Firewalls IDS/IDPS HIPS/HIC/EDR MTD Sandbox CASB/SWG	(Gartner, 2018) (Yadav, 2020) (Bosworth et al., 2014) (Gartner, 2021b)
Information	Databases File Systems E-Mail Systems Web Site Systems	Data Corruption Data Theft	SQL Injection XSS SPAM Dos/DDoS Malware Botnets Password Attacks Phishing	WAF DLP SEG Sandbox DBFW Network Firewalls IDS/IDPS	(Gartner, 2021a) (Yadav, 2020) (Bosworth et al., 2014) (Gartner, 2021b)
Application	Collaboration SW ERP SW CRM SW Knowledge base SW	Applications Disruption	Malware Password Attacks SQL Injection XXS	Network Firewalls IDS/IDPS WAF DLP DBFW	(Yadav, 2020) (Bosworth et al., 2014) (Gartner, 2021b) (Gartner, 2021a)

Table 7 – Asset vs Risks/Threats vs Security Technologies Mapping

4.2.2 Design

In the design phase, organizations shall be conscious of the several types of security controls that need to be implemented, using as guideline the security functional requirements that were described in the theoretical background. For each type of security control, one or more AI technique can be used, which of course have the objective of protecting against the several types of known threats. Moreover, one proposes that in the design phase organizations define a set of KPIs (Key Performance Indicators) to ensure the capability of monitoring, assessing, and evaluating the performance and effectiveness of the chosen solution. This list of KPIs is included in the model for this design phase:

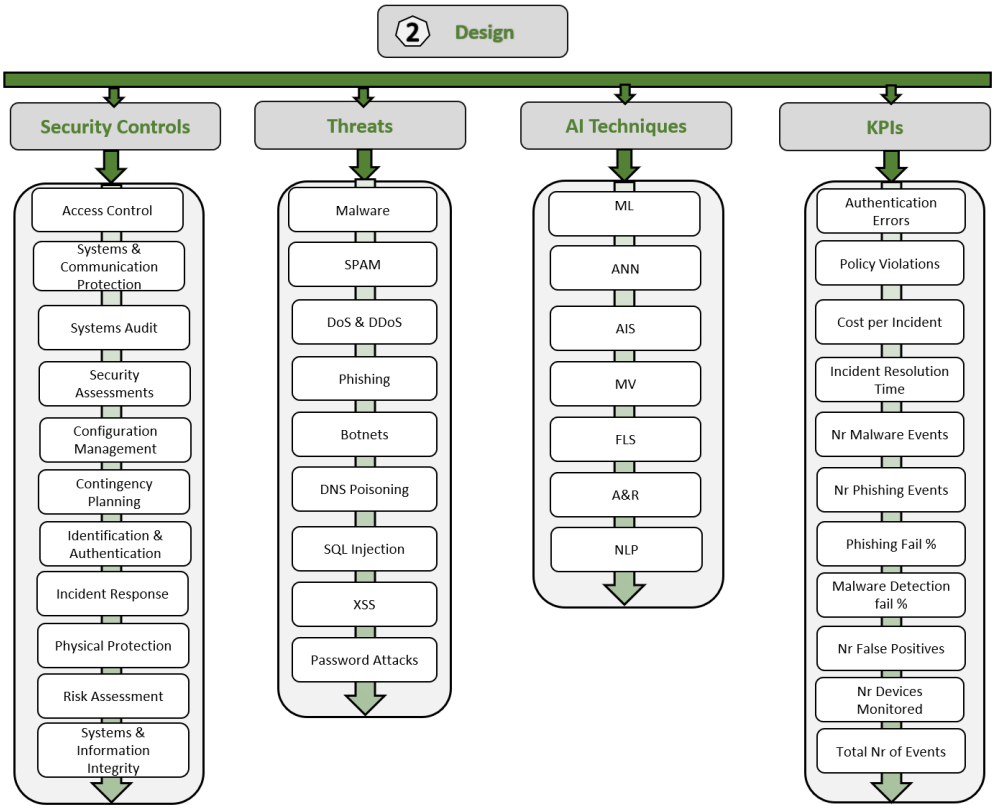

Figure 19 – Proposed Model for the Design Phase

Similarly, to the approach used in the strategy phase, in here it's also proposed a mapping between the various components of the model, in this case, the components are the security controls, AI techniques that can be used to enforce each control, and obviously the several threats that may cause risk to these controls. The following table presents one's proposal for this mapping.

Mapping Matrix

Controls	Security Domain	Threats	AI Technique	References
Access Control Identification & Authentication	Conditional Access Multi Factor Authentication Biometrics Intrusion/Unauthorized Access Prevention Privilege Identity Management	Password Attacks Malware Phishing	DL SVM RF DT NB ANN MV FLS	(Ali et al., 2020) (Aljabri et al., 2021) (Bécue et al., 2021) (Choi et al., 2020) (Lazic, 2019) (NIST, 2006)
Systems & Communications Protection & Integrity	Physical Protection Intrusion Prevention Network Security Traffic Anomaly Information security Data Access Governance Data Encryption	Dos/DDoS Malware SPAM DNS Poisoning Botnets SQL Injection XSS Password Attacks Ransomware Phishing	DL SVM NB RF DT K-Mean DBN ANN AIS MV FLS R&A NLP	(Ali et al., 2020) (Aljabri et al., 2021) (Bécue et al., 2021) (Choi et al., 2020) (Arshi & Madhavi, 2020) (Gbenga et al., 2019) (Lazic, 2019) (NIST, 2006)
Systems Audit	Intrusion Detection Incident Detection Incident Reporting	Malware Ransomware Phishing	DL SVM NB RF DT ANN MV FLS	(Aljabri et al., 2021) (Atiku et al., 2021) (Basit et al., 2021) (Caviglione et al., 2021) (Choi et al., 2020) (Zhao et al., 2002) (NIST, 2006)
Security Assessment	Certification and Compliance	N/A	N/A	(NIST, 2006)
Configuration Management	Asset Management	N/A	N/A	(NIST, 2006)

Controls	Security Domain	Threats	AI Technique	References
Contingency Planning	Business Continuity Disaster Recovery	Ransomware Malware DoS / DDoS	DL SVM NB RF DT ANN MV FLS	(Ali et al., 2020) (Aljabri et al., 2021) (Arshi & Madhavi, 2020) (Choi et al., 2020) (Zhao et al., 2002) (NIST, 2006)
Incident Response	Threat Detection Threat Analysis Threat Recovery	Dos/DDoS Malware SPAM DNS Poisoning Botnets SQL Injection XSS Password Attacks Ransomware Phishing	DL SVM NB RF DT ANN MV FLS NLP	(Ali et al., 2020) (Aljabri et al., 2021) (Arshi & Madhavi, 2020) (Bécue et al., 2021) (Choi et al., 2020) (Gbenga et al., 2019) (H. S. Hussain et al., 2021) (Lazic, 2019) (NIST, 2006)
Physical Protection	Access to Systems, Equipment and Infrastructure Protection Protection from environmental hazards	Physical Intrusion Unauthorized Access Equipment Theft Environmental Disasters	MV R&A NLP	(Horan & Saiedian, 2021) (H. S. Hussain et al., 2021) (Lazic, 2019) (Zhao et al., 2002) (NIST, 2006)
Risk Assessment	Operational Risks Information Systems Risks Data & Information Risks Users Activities Risks Vulnerability Risk Management	Phishing SPAM Password Attacks Ransomware	DL NB MV	(Ali et al., 2020) (Aljabri et al., 2021) (Choi et al., 2020) (Zhao et al., 2002) (NIST, 2006)

Table 8 – Controls vs Domain vs Threats vs AI Techniques Mapping

4.2.3 Implementation

We've seen in the systematic literature review analysis, that some AI based security tools already exist in the market, provided by renown security companies like Check Point, CrowdStrike, FireEye, Fortinet, LogRhythm, Palo Alto Networks, Sophos and Symantec (Lazic, 2019). Knowing this, one proposes that following the strategy and design phases, where the organizations already defined the assets to protect, the technologies required to tackle the threats each asset may face, and to enforce the required security controls, organizations start by scouting the market for existing solutions that could address their needs. If they exist, they should procure it and implement it, if not, then they should consider develop it in house or outsource its development in case that's deemed feasible. Otherwise, they may have to put the AI solution in standby until it becomes available and remain with a legacy solution for the specific case where AI is not possible yet. The implementation itself can either be performed by the solution vendor, or, in case there is in-house expertise, by the local cybersecurity experts. This process can be represented in a simple but objective flowchart like the one in figure 20.

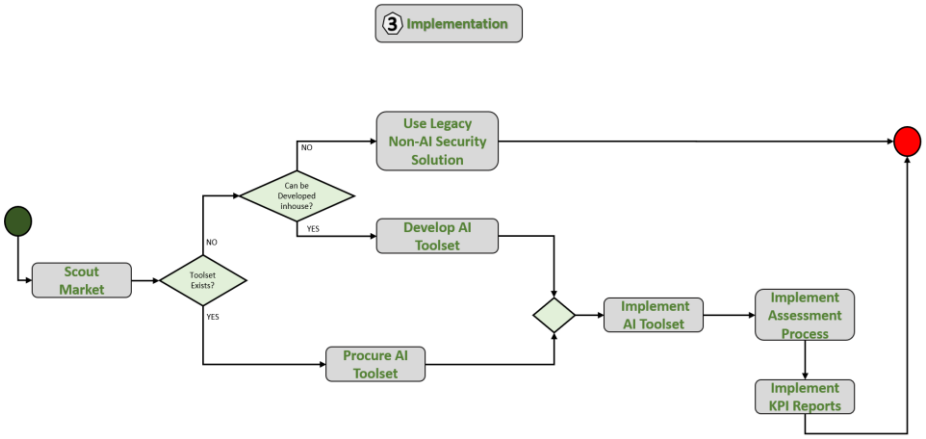

Figure 20 – Proposed Model for the Implementation Phase

As part of this implementation workflow, organizations should also implement an internal process capable of assessing the performance of the AI based security solution, together with a tool capable of reporting the defined KPIs. The assessment of the solution performance comprises of the thorough analysis and evaluation of the reported KPIs, which shall be executed in the operations phase.

4.2.4 Operations

Operations is a core function of any cybersecurity strategy, and this often takes place in what is called the SOC, i.e., Security Operations Centre of an organization. This function is responsible for the permanent monitoring and real time follow up of security incidents. Tasks like incident detection and response, risk and crisis management, security systems log analysis, are core to operations. Also, security events and alerts are collected analysed and acted upon. Security operations is mostly about risk management (Zelonis & Lyness, 2019).

Besides the monitoring aspect, supporting the security systems, troubleshooting issues that may arise, and not least important, perform patch management activities, are paramount for a good operations performance. Based on these principal concepts, the model proposal for the operations phase is represented as follow.

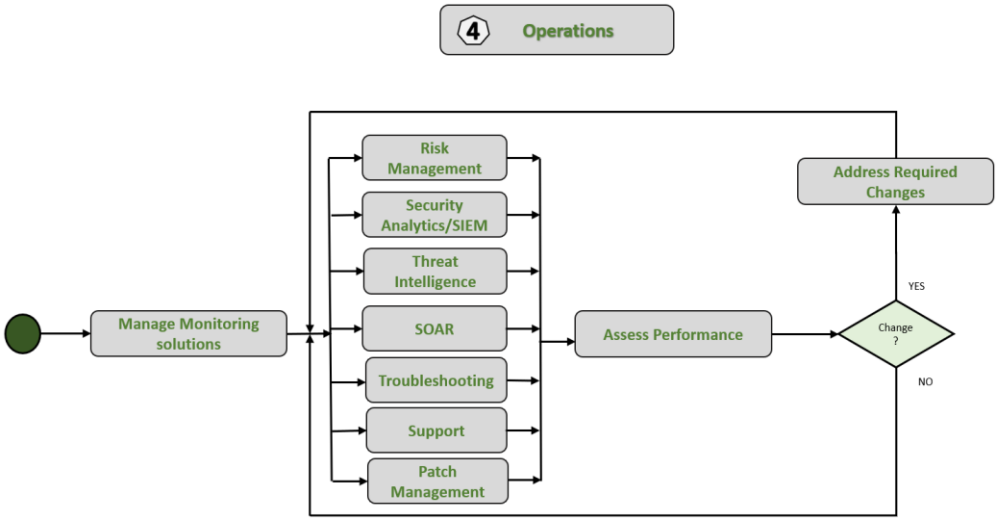

Figure 21 – Proposed Model for the Operations Phase

Besides the core operations activities described, the proposed model also suggests that organizations shall assess the performance of such activities, which in practical terms means executing the assessment process defined in the implementation phase. This assessment shall be performed by constantly testing the technology on: detection, response, recovery, and all other aspects that make part of the operational process. The assessment may also comprise of comparing the KPIs which result of the operations activities, to those defined in the implementation phase. The result of this assessment will define if the solution requires changes and adjustments, in case some KPIs are not met, or some of the functions failed in some testing, or, in case all is according to what's been defined, keep it as it is. When it comes to the execution of the operational model itself, there are a set of activities that shall be used according to operational categories (Valente & Murphy, 2020):

Risk Management

- Risk Assessment and Visibility: which can be achieved using machine identity management tools (ensures confidentiality and integrity of information between machines), continuous control monitoring tools and vulnerability control tools.
- Risk Quantification: Provide security ratings for the potential threats and vulnerabilities, in the form of CVSS (Common Vulnerability Scoring System) scores.
- Penetration Testing Activities: Simulation of real attacks using publicly available databases of known exploits, allowing the security teams to be trained in detecting and stopping attacks when they are real.
- Breach Simulation Activities: Provides the security teams with the ability to test thousands of adversarial attacks in a short timeframe which target several types of systems, allowing them to understand the impact in may have once faced with the real thing.

Security Analytics / Security Information and Event Management (SIEM)

- Centralized collection of events logs from several security systems, and correlation of these events in real-time
- Real-time analysis of events data for early detection of attacks and breaches
- Collection, storage, investigation, and reporting of events data for incident response, forensics, and regulatory compliance
- Analysis and visibility of activities inside the network
- User behaviour analytics, i.e., gathering of insights related to users' activities to identify malicious users and accounts that may have been compromised.

Cyber Threat Intelligence

- Usage of shared information between security entities, that provides mechanisms to identify and prevent attacks, recognize adversaries, malware and other already known threats.
- Intelligence is provided by several sources of information agencies, and security products can source this information directly from feeds.

- Types of intelligence comprises of social media intelligence, dark web intelligence, brand risk analytics.

Security Orchestration Automation and Response (SOAR)

- Coordination, execution, and automation of tasks between various people in the organization and security tools, therefore providing mechanisms for organizations to respond quickly to cyber-attacks, improving their overall security posture.

Patch Management

- Patching is the practise of updating systems software or any type of code, to address vulnerabilities that could be exploited by attackers. Patching is a crucial activity in cybersecurity operations, or any type of IT operations

The operational categories described and its correspondent activities can also be achived with AI technologies, according to the matrix in the following table (Bhandari, 2021), (Gray, 2020).

Mapping Matrix						
	AI Techniques					
Operational Categories	ML	ANN	MV	R&A	FLS	NLP
Risk Management	✓	✓	✓	✓	✓	✓
Security Analytics	✓	✓			✓	✓
Threat Intelligence	✓	✓			✓	
SOAR	✓	✓		✓		
Patch Management	✓			✓		

Table 9 – Operational Categories vs AI Techniques Mapping

4.2.5 Roadmap

The 5th and final stage of the proposed framework is called the roadmap phase. In this phase organizations shall keep a mindset of constant evaluation of the entire framework and its effectiveness in their overall security strategy, with the objective of maintaining a continuous improvement process. At this point what's evaluated is if what's been done so far is providing results according to the strategy defined in phase one, and if the KPIs in use are the necessary and the most relevant to fulfil this strategy. This is not like in the operations phase, where what's being assessed is the effectiveness of the solution, i.e., if it's performing well, and if the defined KPIs are met or not. At this stage it's all about understanding if both the KPIs in use and the implemented solution are the right ones to meet the needs of the organization.

We can represent this workflow with a model like the one in figure 22.

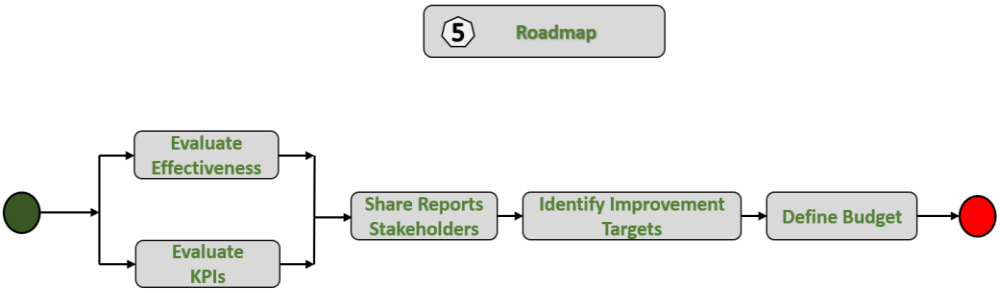

Figure 22 – Proposed Model for the Roadmap Phase

The workflow basically comprises of the evaluation of the KPIs and the solution effectiveness, as described before, followed by sharing reports with identified stakeholders for analysis and decision making. Once analyzed, improvement targets may be identified, and if so, the organization shall define a budget to be applied in the improvement of the overall security strategy.

4.2.6 Solution Architecture

Implementation of the proposed framework involves several aspects, but a core aspect to all this is of course the technology used. When using technology, it's a good approach to represent the several components and its interactions in a diagram representing the solution architecture. One's proposal for a possible solution architecture which addresses this security framework is presented in figure 23.

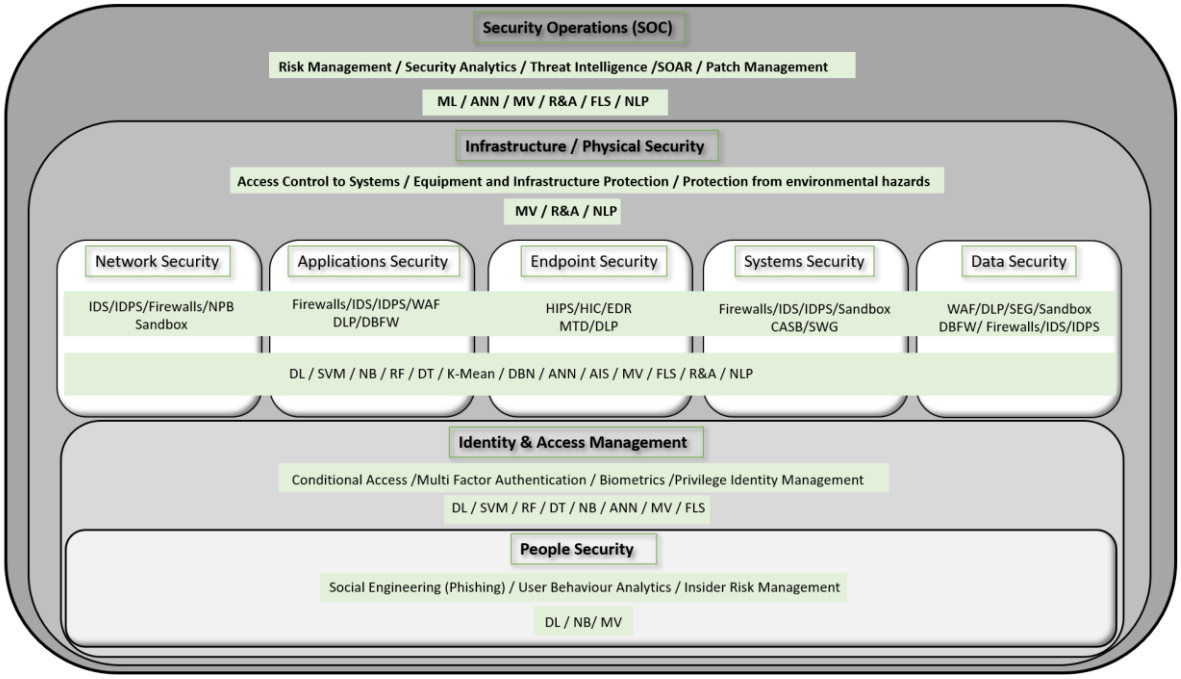

Figure 23 – Solution Architecture

4.2.7 Validation & Discussion

Having developed the framework, i.e., the artifact of one's research, we reach the 4th stage of the DSRM model, the evaluation stage of the design science research strategy. As described in section 2.2, one used the Peffers model for the DSR strategy and the evaluation method chosen was the collection of feedback from subject matter experts (Peffers et al., 2007). The feedback was collected by the means of interviews performed using online meetings, with the purpose of validating the proposed framework, aiming to receive critical comments and suggestions for improvements, where deemed justifiable, therefore filling some of identified research gaps. One has chosen experts from the industry, i.e., cybersecurity professionals, and from the academic area, because in each of these areas there are different views, different levels of experience and expertise that in the end are complementary. In these meetings, before reaching the interview phase itself where a set of defined questions were posed to the experts, one has done a brief presentation of the framework using a slide pack which started with the background and objectives of the study as a baseline for the framework understanding. The following table presents the list of experts, their professional role and are of expertise.

#	Professional Role	Area of Expertise	Domain
E1	Global Cybersecurity Designer @ Vodafone Cybersecurity Organization	Cybersecurity Professional	Industry
E2	Assistant Professor @ Nova IMS	Professorship	Academic
E3	Cybersecurity Trainer & Consultant @ Centro Nacional de Cibersegurança	Cybersecurity Professional	Industry
E4	Cybersecurity Information Consultant @ ORAMIX	Cybersecurity Professional	Industry

Table 10 – Experts Interviewed

These meetings have been recorded, and both the feedback and the answers to the interview questions have been written down for later analysis. The discussion of this feedback is presented next, and the full transcript of answers to the interview's questions are presented in annexes of this thesis document.

A common and very positive reaction received from the four experts interviewed is that they all say the framework is very complete, presented in a clear and detailed way, and technically accurate when it comes to address the main aspects of cybersecurity threats and technologies. The technical aspect of it and the detail of the information presented was also praised. Equally they all had a similar feedback when it comes to comment the AI utilization as it's suggested in the framework, i.e., being people from the "field" they struggle to find a direct link between what the major security vendors currently provide in terms of AI based tools, and the AI utilization proposed in the framework, which of course is based on a deep study of available scientific documentation, and not industry or vendor specific solutions. None less there was a consensus that, although what they know from the field is mostly based on ML, the proposed applicability of other AI techniques make sense, and they believe we may start to see something in the near future.

The five main steps used for the high-level framework also received positive feedback from all the experts. Despite the overall positive feedback, all of them have contributed with some valid inputs which are considered to bring additional value to the work. Experts 1 and 2 were more detailed and

provided, in one's opinion, the most valuable inputs. Experts 3 and 4 were not as critical, but still the comments received were considered equally valuable and therefore taken into consideration for the revised framework.

Expert 1 is a very technical person, and his comments are therefore very pertinent on the technical side. First suggestion is regarding the design stage where he considers that the training of AI technology is an important activity to include in this stage. In the implementation phase, besides having the activity of "Implement an assessment process of the solution", this expert believes it should also be included an activity for "implementing an assessment process for the AI automated output". In the design phase, where one has described the threats applicable to "Access Control, Identification & Authentication" security controls, expert 1 states that only password attacks should be considered, and malware and phishing attacks shouldn't be part of it. Another important input is related to the use of "Conditional Access" as a security domain in these security controls, where he states that this is a vendor specific term, and it should probably be replaced with a "zero trust implementation" mechanism type. On the technology's aspect, expert 1 pointed out that sandbox is a capability and not a technology from his point of view, in the sense that, looking from the framework implementation perspective, the AI toolset would detect a threat and put in in a sandbox for analysis. It's the AI technology that triggers the security activity, so he recommends that one should probably remove the sandbox from the security technologies list. A very pertinent observation, in one's opinion, is that the framework is missing a reference for encryption technologies, which of course are a core part of cyber security and as such, it was suggested that encryption technologies are added to all asset types. Final observation from expert 1 has to do with the security architecture diagram. The diagram is a good representation; however, he doesn't see it as a "Security Architecture" like one has called it, but a more appropriate naming could be "Stages of the AI Architecture". Following the criticism and recommendations for improvement described, when asked about the utility of this framework, expert 1 said that the framework is useful from an overall high-level concept of the implementation of AI technology, and it could be used as a starting point for the discussion of implementation of AI in the organization. When asked about if he would consider implementing the framework, expert 1 said that if a business was to implement an AI based security solution, this framework would be good to provide a scope of where in the business AI could be implemented and which AI techniques could be used, which threats they would be facing and risk reduction it could be achieved by implementing it. Since it's also provided a list of security controls, mapping to the threats and applicable AI technologies, then this would provide a good overview of risk reduction.

Expert 2 brings a mix of industry experience and academical background, which also resulted in a very rich insight. He started by addressing the way one has defined the cybersecurity assets and how they have been used in the framework. He pointed out that there is an industry rfc (rfc 4949) that describes the security asset types and how they're classified, and in this sense, he stated that what is presented in the framework is very identical to what the rfc defines, therefore very good. From his experience and field knowledge, the AI techniques made available by the security vendors are mostly based on machine learning. While discussing this point one has highlighted that the result of the proposal is based on the output of scientific papers where there is a lot of focus on specific ML classifiers and techniques that are in essence a subset, or type, of machine learning. On this, the expert suggested that perhaps the vendors do not provide further detail on the specific techniques used either due to marketing reasons, or perhaps in some cases, due to intellectual property reasons

and put it all under the same “umbrella” of machine learning. In any case, when looking at one’s proposal this expert sees applicability of other AI techniques proposed in the framework. He pointed out that where one described “People Security”, he sees applicability of NLP. As an example, we could use NLP to detect phone calls received by people and detect if it’s a human call or a bot, that is, make use of this AI technique to protect against scams that use automated phone calls. As a fact, expert 2 has knowledge of some NLP applications being used by some nation’s secret services, which perform language analysis, and this can be used to detect phishing scams. When it comes to secret services, we’re taking about filtering millions of calls, so using NLP here, can help identify which calls are relevant to listen to its content and which calls should be discarded. If we transpose this to an organization security, they probably will have some privacy concerns to address, which secret services don’t need to worry with, however the technical capability of utilizing this AI technique in cybersecurity is viable. Going back to the subject of the applicability of AI techniques in cybersecurity real cases, expert 2 states that his experience is mainly of ML usage, any others are not much used yet. Inside ML e states that supervised learning is the most used technique, and this is because in different types of businesses the system must be trained in different ways. As an example, a telco company will be targeted with a type of SPAM email which is different than a bank, resulting in different training algorithms that are required to perform the email reputation. When asked about the utility of this framework, Expert 2 considers it useful, but he would like to understand how this could be used to detect and prevent critical incidents, or reduce false positives rate, because these are the type of metrics that are relevant from a SOC perspective and have implications in terms of risk management and productivity levels. One argued that we can see in the framework which AI techniques can be used in different types of the main SOC activities, however it’s agreed that the framework does not enter in such detailed analysis. In terms of main criticism, he stated that there not much to point out, except perhaps for the fact that it uses a set of AI techniques that probably are too technical, but at the same time he recognizes that this is how a framework is, so being this a very broad and complex topic, it would be difficult to present a narrower scope. At this point he emphasizes again the fact that AI utilization in the field of cybersecurity is mostly focused on machine learning, and other AI techniques are still very much addressed at an academical level. When asked about if he would use this framework, he starts by putting the use of a framework in the context of having a specific objective in terms of security. Here we must consider that we have a set of assets we need to protect, and for that, organizations must perform an investment. In the end organizations wish to reduce risk and with this in mind he believes that probably the framework can be useful because AI mechanisms are more effective, therefore can contribute better for risk reduction. However, when it comes to the decision-making process to use AI based techniques, organizations must do a bigger investment, because this normally implies the subscription of a premium service. At this point it’s important to evaluate if the investment in a premium service makes sense, when comparing the investment to the value of the assets being protected and the risks, they’re subject to. Having said this, another suggestion came up, which is related to the budgeting exercise. In the framework one has suggested to define budget for improvement in the roadmap phase, but obviously that in the strategy phase, we must also define the initial budget for the solution to implement. This budget shall come out from the risk evaluation vs asset value exercise. In the end, and as a conclusion, expert 2 states that his main recommendation for improvement would be exactly this budget definition in the strategy phase.

Expert 3 brings experience from the consultancy field together with some training experience. First comment received was regarding the assets type definition and the mapping of threats and risks to each asset type. Here this expert pointed out that in "Centro Nacional de Cibersegurança" (CNCS) a very similar definition is used, and this was praised as a very positive point. He continues by stating that the high-level framework is very in line with the risk management framework used by CNCS, but this one has the added value introduced by the utilization of AI. Still evaluating the high-level framework, expert 3 pointed out that the strategy phase is also very in line with CNCS but would be better if it could include a risk evaluation matrix, which has the purpose to define the risk level for the business posed by the several type of threats. This matrix would be a combination of the probability of a certain threat occur and how this treat is seen in terms of damage inflicted to the business. Basically, we're talking about evaluating the business impact caused by the occurrence of a certain threat classifying it as low, medium, or high impact. One's analysis on this feedback takes us back to the suggestion expert 2 has provided related to the budgeting exercise. In that case, it was suggested that a similar type of risk evaluation should be done, which then would help define the investment budget for a security solution of this type. Expert 3 also noticed that in the roadmap phase one is proposing that the evaluation of the solution effectiveness and the KPI reporting are done in parallel. From his point of view this either should be consolidated in a wider single activity, or just be performed in a sequential way. When asked if the framework is useful, expert 3 said yes because it can be a complementary implementation in relation to the Portuguese National Cybersecurity Framework defined by Portuguese National Cybersecurity Centre, and with the utilization of the AI based techniques of the proposed framework, the response using threat intelligence would step up to a next level of security. When then asked about if he would implement this framework, he also said yes because with the necessary adaptation to Portuguese National Cybersecurity Framework, and some further development to perform a case study, it could be possible to apply it in most of the public government infrastructure.

Feedback from a 4th expert, which also comes from the cybersecurity consultancy field is not as detailed as expert 1 and 2 but also provided valuable insights. Expert 4 stated that the framework is technically very good, and there is nothing critical to point out. However, similarly to expert 2 feedback, he said that the reference to so many AI techniques are not seen in the current industry offering which are very much focused on machine learning. None less, the overall proposal makes sense and it's reasonable to expect that we start seeing something new in this space soon. Expert 4 also highlighted the importance of having a budget definition activity based on a risk evaluation process in the strategy phase, which pretty much aligns with what's been said before. The budgeting exercise proposed for the roadmap phase is correct and makes sense, but it's paramount that it's also considered in the strategy phase aligned with the risk evaluation. Identifying the risks, which is already proposed, but also evaluating its business impact and suggesting risk treatment measures is something that organization should do in the strategy phase which helps define the initial budget. Finally, when asked if the framework is useful, expert 4 said yes because it is well designed and it covers all the technologies and risks associated with them. Then, when asked if he would use the framework he also said yes because this framework would help the operations to be smother and less stressful to the security teams.

4.2.8 Revised Framework

The result of the interviews with the 4 experts resulted in a rich set of opinions, which were mostly positive, but also some criticism that must be taken as suggestions for improvement. The discussion addressed in the previous section has several aspects depending on the expertise and experience of the different people. From all the recommendations for improvement one believes that what is of most relevance and has the most impact in the effectiveness of its implementation are the references related to creation of a risk evaluation matrix along with the initial budget definition, which was common between the 4 experts. The suggestion to perform the AI technologies training in the design stage and the assessment of the AI output in the implementation phase suggested by expert 1 also deserves to be considered as part of the high-level framework improvement. The result of it is presented in the following picture.

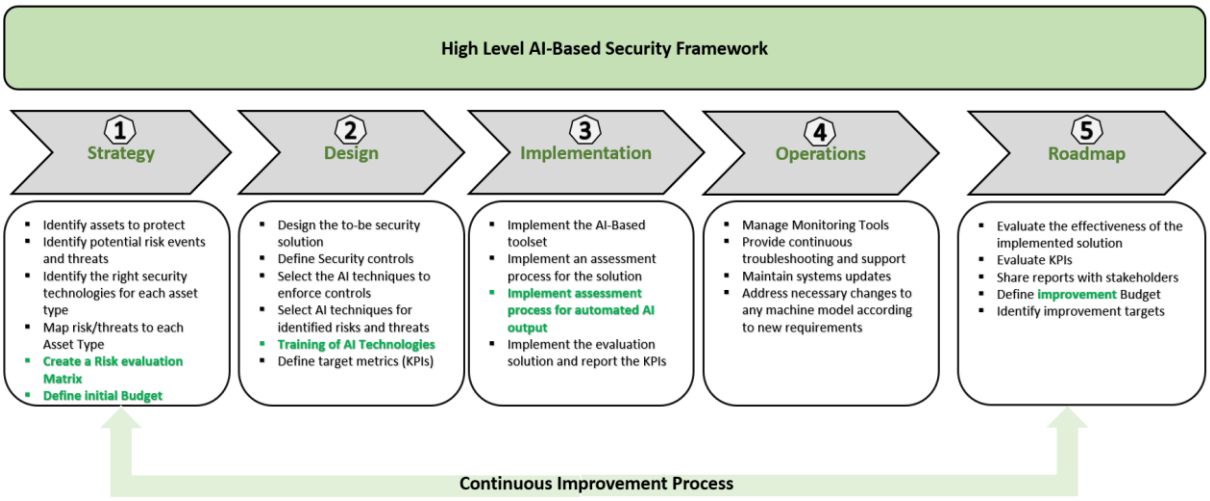

Figure 24 – Revised High-Level Framework

Going deep into the detailed framework stages, one has taken as most relevant suggestion the recommendation provided by expert 1 related to the inclusion of encryption technologies as part of all the asset types. This results in the following detailed diagram.

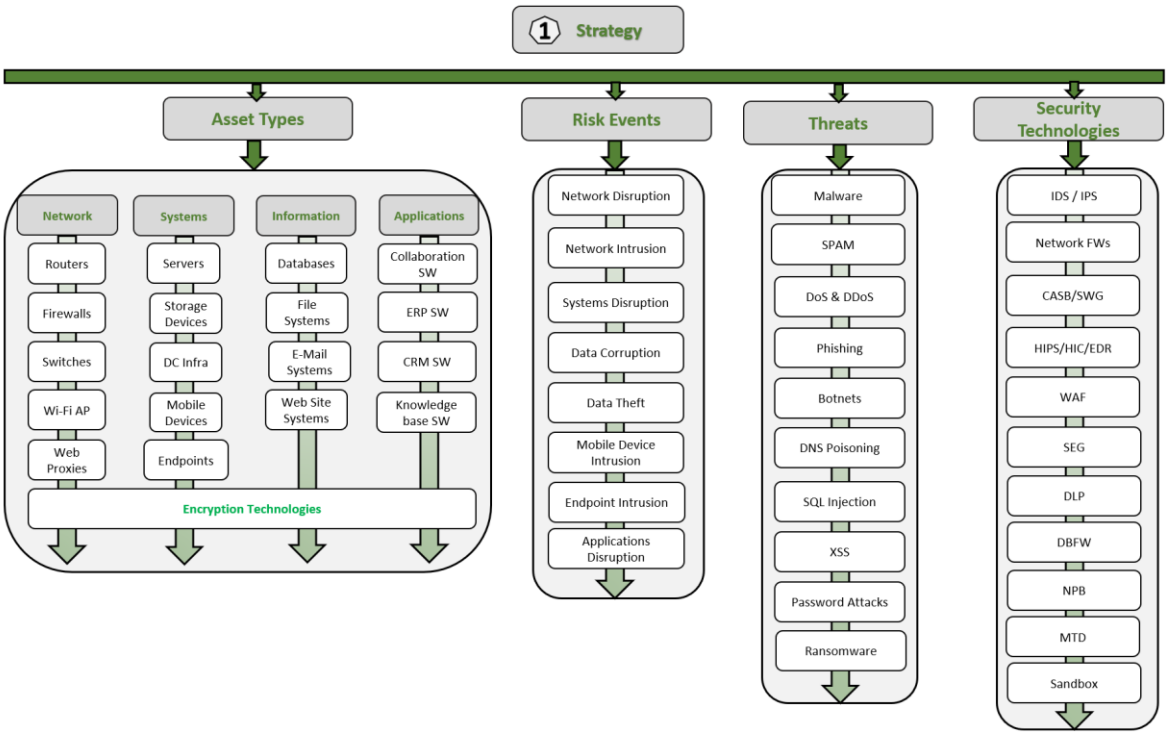

Figure 25 – Revised Framework – Strategy Phase

Not least relevant from one's point of view, is the alert for the wrong utilization of malware and phishing as threats posed to "access control, identification & authentication" security controls. Also of critical importance is the removal of the reference for "Conditional Access" which is a vendor specific naming and not an industry standard. This is reflected in a revised mapping matrix for the design phase as presented here.

Mapping Matrix

Controls	Security Domain	Threats	AI Technique	References
Access Control Identification & Authentication	~~Conditional Access~~ Multi Factor Authentication Biometrics Intrusion/Unauthorized Access Prevention Privilege Identity Management	Password Attacks ~~Malware~~ ~~Phishing~~	DL SVM RF DT NB ANN MV FLS	(Ali et al., 2020) (Aljabri et al., 2021) (Bécue et al., 2021) (Choi et al., 2020) (Lazic, 2019) (NIST, 2006)
Systems & Communications Protection & Integrity	Physical Protection Intrusion Prevention Network Security Traffic Anomaly Information security Data Access Governance Data Encryption	Dos/DDoS Malware SPAM DNS Poisoning Botnets SQL Injection XSS Password Attacks Ransomware Phishing	DL SVM NB RF DT K-Mean DBN ANN AIS MV FLS R&A NLP	(Ali et al., 2020) (Aljabri et al., 2021) (Bécue et al., 2021) (Choi et al., 2020) (Arshi & Madhavi, 2020) (Gbenga et al., 2019) (Lazic, 2019) (NIST, 2006)
Systems Audit	Intrusion Detection Incident Detection Incident Reporting	Malware Ransomware Phishing	DL SVM NB RF DT ANN MV FLS	(Aljabri et al., 2021) (Atiku et al., 2021) (Basit et al., 2021) (Caviglione et al., 2021) (Choi et al., 2020) (Zhao et al., 2002) (NIST, 2006)
Security Assessment	Certification and Compliance	N/A	N/A	(NIST, 2006)
Configuration Management	Asset Management	N/A	N/A	(NIST, 2006)

Controls	Security Domain	Threats	AI Technique	References
Contingency Planning	Business Continuity Disaster Recovery	Ransomware Malware DoS / DDoS	DL SVM NB RF DT ANN MV FLS	(Ali et al., 2020) (Aljabri et al., 2021) (Arshi & Madhavi, 2020) (Choi et al., 2020) (Zhao et al., 2002) (NIST, 2006)
Incident Response	Threat Detection Threat Analysis Threat Recovery	Dos/DDoS Malware SPAM DNS Poisoning Botnets SQL Injection XSS Password Attacks Ransomware Phishing	DL SVM NB RF DT ANN MV FLS NLP	(Ali et al., 2020) (Aljabri et al., 2021) (Arshi & Madhavi, 2020) (Bécue et al., 2021) (Choi et al., 2020) (Gbenga et al., 2019) (H. S. Hussain et al., 2021) (Lazic, 2019) (NIST, 2006)
Physical Protection	Access to Systems, Equipment and Infrastructure Protection Protection from environmental hazards	Physical Intrusion Unauthorized Access Equipment Theft Environmental Disasters	MV R&A NLP	(Horan & Saiedian, 2021) (H. S. Hussain et al., 2021) (Lazic, 2019) (Zhao et al., 2002) (NIST, 2006)
Risk Assessment	Operational Risks Information Systems Risks Data & Information Risks Users Activities Risks Vulnerability Risk Management	Phishing SPAM Password Attacks Ransomware	DL NB MV	(Ali et al., 2020) (Aljabri et al., 2021) (Choi et al., 2020) (Zhao et al., 2002) (NIST, 2006)

Figure 26 – Revised Mapping Matrix – Design Phase

5 CONCLUSIONS

After performing an extensive literature review, firstly to consolidate knowledge on the concepts of cybersecurity and artificial intelligence, followed by a systematic literature review to understand the state of the art related to the utilization of AI in cybersecurity, what are the main challenges faced by organizations in this area and which are the advantages and disadvantages of the utilization of AI in cybersecurity, it was possible to conclude that although there is a lot of scientific literature focused on this topic, several studies, developed models and security ontologies, there is still a gap to fill in this space. However, it was also possible to conclude that the existing AI technologies are evolving in the direction of making it possible to provide security solutions based on AI that are robust and more effective against AI based cyber-attacks, when compared to the traditional solutions. Additionally, it was possible to conclude that there is already a growing interest from most organizations in the adoption of this type of more advanced cybersecurity solutions. In this sense it could also be concluded that the framework one proposed as the objective of this work, was viable to be developed, has its space in the cybersecurity landscape and is pertinent. The discussion and analysis held with several experts, from the industry and the academic domain, proved that, despite some criticism which was mostly constructive and resulted in suggestions for improvement, the work developed achieved its objective.

5.1 SYNTHESIS OF THE WORK DEVELOPED

The developed work followed a structured approach. Firstly, one has performed a background analysis to identify the problem of the studied topic, understand its relevance and importance and then define its objective. Following this initial approach, and having defined the objective, one needed to acquire a background knowledge on the two main topics in study, and for this purpose one conducted a literature review with the purpose of building a solid theoretical background knowledge on cybersecurity and artificial intelligence. The next step was then to perform a systematic literature review focused on the utilization of AI in the cybersecurity domain, with the purpose of finding the answers for a set of research questions. These answers formed the foundation which allowed the creation and development of the framework. This framework which one expects to bring some benefit for the industry and the academic world is the final objective of the work. Finally, having developed the framework, and in line with the research strategy defined, one has performed several presentations to experts in the area to collect feedback, criticism, and recommendations for improvement. The outcome of the discussion with the several experts resulted in a revised framework where some of their recommendations were adopted.

5.2 RESEARCH LIMITATIONS

The evaluation and validation of the proposed framework has some limitations. Firstly, although all the experts interviewed are very knowledgeable and experienced people who provided positive feedback and contributed to some improvements, it was only possible to interview four experts. It would have been more beneficial to have the framework evaluated by a wider spectrum of

experienced professionals. The experts interviewed claim that major vendors only offer AI based security solutions which rely on machine learning, however the scientific documentation shows us a much wider range of AI techniques used in this domain. Since this is an academic study, it wasn't possible to argue with the experts on this. In this sense it would have been interesting if one had the chance to speak to someone from the major security vendors who claim to provide AI based security solutions, and cross validate the proposed solution to what the industry is currently offering.

Also, this evaluation was only performed from a theoretical perspective, and it wasn't possible to implement it in a real-life situation. Since the objective of the framework is to be utilized by organizations, to have a clear conclusion about its effectiveness, it's essential to test it in a practical and real situation. Only this way one can have a more realistic view about what works and what else needs to be improved.

Moreover, due to time limitation, it wasn't possible to perform the communication stage as described in the research strategy.

5.3 FUTURE WORK

As for the future work it stands out as an obvious important activity the implementation of this framework in a real situation inside an organization. This will allow to perceive its real effectiveness and collect tangible results. The proposed framework suggests the utilization of some AI techniques which are seen by the experts as something more academical, despite the literature suggesting otherwise. This is a topic that certainly can be clarified by a real-life applicability of the framework. On top of this, it emerges the need to keep up with the industry trends in this space and follow up the evolution and development of technology using these AI methodologies. For some organizations, it might be relevant to assess the possibility to develop some in-house AI based solutions where the vendors may still be missing some offer. Based on the conclusions and results of a real applicability, where gaps are found, the future work must surely focus on improvements and adaptations of the current proposal.

The last stage of the research strategy used in this work, i.e., the communication phase, should also be considered as a proposal for future work. Making this study available for a wider academic community, via some renown publication, would possibly be of utility for other researchers working on these subjects, help them on their own research or trigger their interest towards some other relevant investigations.

BIBLIOGRAPHICAL REFERENCES

Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. A. (2019). Applications of artificial intelligence in transport: An overview. *Sustainability (Switzerland)*, *11*(1). https://doi.org/10.3390/su11010189

Abraham, A. (2005). NEURAL NETWORKS. In P. H. Sydenham & R. Thorn (Eds.), *Handbook of Measuring System Design*. Sons, John Wiley &.

Alharbi, A., Seh, A. H., Alosaimi, W., Alyami, H., Agrawal, A., & Kumar, R. (2021). Analyzing the Impact of Cyber Security Related Attributes for Intrusion Detection Systems. *Sustainability (Switzerland)*, 1–19.

Ali, R., Ali, A., & Aleem, S. (2020). *A Systematic Review of Artificial Intelligence and Machine Learning Techniques for Cyber Security*. *August*. https://doi.org/10.1007/978-981-15-7530-3

Aljabri, M., Aljameel, S. S., Mohammad, R. M. A., Almotiri, S. H., Mirza, S., Anis, F. M., Aboulnour, M., Alomari, D. M., Alhamed, D. H., & Altamimi, H. S. (2021). Intelligent Techniques for Detecting Network Attacks: Review and Research Directions. *Sensors MDPI*, 1–43.

Alwaghid, A. F., & Sarkar, N. I. (2020). Exploring Malware Behavior of Webpages Using Machine Learning Technique : An Empirical Study. *Electronics*, 1–20.

Amit. (2016). *Understanding the difference between cybersecurity and information security*. CISO Platform. https://www.cisoplatform.com/profiles/blogs/understanding-difference-between-cyber-security-information

Amoroso, E. (2006). Cyber Security. *New Jersey: Silicon Press*.

Arshi, M., & Madhavi, K. (2020). A Survey of DDOS Attacks Using Machine Learning Techniques. *E3S Web of Conferences ICMED 2020*, *01052*, 1–6.

Aslan, Ö. (2020). A Comprehensive Review on Malware Detection Approaches. *IEEE Access*, *8*, 6249–6271. https://doi.org/10.1109/ACCESS.2019.2963724

Atiku, S. B., Aaron, A. U., Job, G. K., Shittu, F., & Yakubu, I. Z. (2021). Survey On The Applications Of Artificial Intelligence In Cyber Security. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *October*.

Barraclough, P. A., Fehringer, G., & Woodward, J. (2021). Intelligent cyber-phishing detection for online. *Computers & Security*, *104*, 102123. https://doi.org/10.1016/j.cose.2020.102123

Barrett, Matt, N. (2018). Framework for improving critical infrastructure cybersecurity. *Proceedings of the Annual ISA Analysis Division Symposium*, *535*, 9–25.

Basit, A., Zafar, M., Liu, X., Rehman, A., Zunera, J., & Kashif, J. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, *76*(1), 139–154. https://doi.org/10.1007/s11235-020-00733-2

Bécue, A., Praça, I., & Gama, J. (2021). Artificial intelligence, cyber-threats and Industry 4.0. *Artificial Intelligence Review*, *54*(5), 3849–3886. https://doi.org/10.1007/s10462-020-09942-2

Bello, I., Chiroma, H., Abdullahi, U. A., Ya, A., & Jauro, F. (2021). Detecting ransomware attacks using intelligent algorithms : recent development and next direction from deep learning and big data perspectives. *Journal of Ambient Intelligence and Humanized Computing*, *12*(9), 8699–8717. https://doi.org/10.1007/s12652-020-02630-7

Bhandari, P. (2021). *Enabling AI-powered Smarter Cybersecurity Solutions*. Continuous Security. https://www.xenonstack.com/artificial-intelligence-solutions/cyber-security/

Bose, B. K. (2017). Artificial Intelligence Techniques in Smart Grid and Renewable Energy Systems - Some Example Applications. *Proceedings of the IEEE*, *105*(11), 2262–2273. https://doi.org/10.1109/JPROC.2017.2756596

Bosworth, S., Kabay, M. E., & Whyne, E. (2009). Handbook Computer Security. In *John Wiley & Sons, Inc* (5th ed.). John Wiley & Sons, Inc. http://gso.gbv.de/DB=2.1/PPNSET?PPN=595811353

Bosworth, S., Whyne, E., & Kabay, M. E. (2014). *Computer Security Handbook* (S. Bosworth, E. Whyne, & M. E. Kabay (eds.); 6th ed.). Wiley.

Braitenberg, V. (1998). Vehicles Experiment in Synthectic Psychology. In *MIT Press* (Vol. 4, Issue 4, p. 1). MIT Press. http://www.ncbi.nlm.nih.gov/pubmed/16857884%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1838502&tool=pmcentrez&rendertype=abstract%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3590906&tool=pmcentrez&rendertype=abstract%5Cn

Canongia, C ; Mandarino, R. (2014). Cybersecurity: The New Challenge of the Information Society. In Crisis Management: Concepts, Methodologies, Tools and Applications. In *Hershey, PA: IGI Global*. http://dx.doi.org/10.4018/978-1-4666-4707-7.ch003

Cascavilla, G., & Tamburri, D. A. (2021). Cybercrime threat intelligence : A systematic multi-vocal literature review. *Computers & Security*, *105*, 102258. https://doi.org/10.1016/j.cose.2021.102258

Castro, L. N. De, & Timmis, J. (2002). An Artificial Immune Network for Multimodal Function Optimization. *Proceedings of the 2002 Congress on Evolutionary Computation. IEEE*, 699–704.

Caviglione, L., Choraś, M., Corona, I., & Member, S. (2021). Tight Arms Race : Overview of Current Malware Threats and Trends in Their Detection. *IEEE Access*, 5371–5396. https://doi.org/10.1109/ACCESS.2020.3048319

CCRA. (1996). *The Common Criteria*. 1996. https://www.commoncriteriaportal.org/

Chan, L., Morgan, I., Simon, H., Alshabanat, F., Ober, D., Gentry, J., Min, D., & Cao, R. (2019). Survey of AI in cybersecurity for information technology management. *2019 IEEE Technology and Engineering Management Conference, TEMSCON 2019*. https://doi.org/10.1109/TEMSCON.2019.8813605

Chandra, S., Paira, S., Alam, S. S., & Sanyal, G. (2014). A comparative survey of symmetric and asymmetric key cryptography. *International Conference on Electronics, Communication and Computational Engineering*, 83–93.

Choi, Y., Liu, P., Shang, Z., Wang, H., Wang, Z., Zhang, L., & Zhou, J. (2020). Using deep learning to solve computer security challenges : a survey. *Springer Open*.

CNSS. (2010). National Information Assurance (IA) glossary. *The National Security Systems Instruction*, *4009*, 103. http://www.cnss.gov/Assets/pdf/cnssi_4009.pdf

Conteh, N. Y., & Schmick, P. J. (2016). Cybersecurity : risks , vulnerabilities and countermeasures to prevent social engineering attacks. *International Journal of Advanced Computer Research*, *6*(23).

Cord, M., & Delany, S. J. (2008). Supervised Learning. In Springer (Ed.), *Machine learning techniques*

for multimedia. Springer International Publishing.

Costa, E., & Simões, A. (2008). *Inteligência Artificial Fundamentos e Aplicações* (F.-E. de Informática (ed.); 3rd ed.). FCA - Editora de Informática.

Craigen, D., Diakun-Thibault, N., & Purse, R. (2014). Defining Cybersecurity. *Technology Innovation Management Review*, *4*(10), 13–21. https://doi.org/10.22215/timreview835

Dayalan, M. (2020). Cryptography in Computer Security. *JETIR*, *May 2019*. https://doi.org/10.1717/JETIR.17140

Devi.T, R., & Pradesh, A. (2013). Importance of Cryptography in Network Security. *IEEE, Computer Society*, 462–467. https://doi.org/10.1109/CSNT.2013.102

DHS. (2014). A Glossary of Common Cybersecurity Terminology. In *National Initiative for Cybersecurity Careers and Studies: Department of Homeland SecurityInitiative for Cybersecurity Careers and Studies: Department of Homeland Security*. http://niccs.us-cert.gov/glossary#letter_c

DoD. (1985). *TRUSTED COMPUTER SYSTEM EVALUATION CRITERIA*. https://csrc.nist.gov/csrc/media/publications/conference-paper/1998/10/08/proceedings-of-the-21st-nissc-1998/documents/early-cs-papers/dod85.pdf

Dutt, I. (2020). Immune System Based Intrusion Detection System (IS-IDS): A Proposed Model. *IEEE*.

Farzadnia, E., Shirazi, H., & Nowroozi, A. (2021). A novel sophisticated hybrid method for intrusion detection using the artificial immune system. *Journal of Information Security and Applications*, *58*(February), 102721. https://doi.org/10.1016/j.jisa.2020.102721

Feng, X. (2020). Artificial Intelligence Cyber Security Strategy. *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing*, 328–333. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00064

Fine, T. L. (2016). Fundamentals of Artificial Neural Networks. *IEEE Transactions on Information Theory*, *August 1996*, 2–5. https://doi.org/10.1109/TIT.1996.508868

Flores-fuentes, W., Rodríguez-quiñonez, J. C., Hernandez-balbuena, D., Rivas-lópez, M., Sergiyenko, O., Felix, F., & Rivera-castillo, J. (2014). *Machine Vision supported by Artificial Intelligence Applied to Rotatory Mirror Scanners*. 1949–1954. https://doi.org/10.1109/ISIE.2014.6864914

Franklin, S., & Graesser, A. (1997). *A Software Agent Model of Consciousness*.

Gamage, S., & Samarabandu, J. (2020). Deep learning methods in network intrusion detection : A survey and an objective comparison. *Journal of Network and Computer Applications*, *169*(February), 102767. https://doi.org/10.1016/j.jnca.2020.102767

Gartner. (2018). *TrustSpace ; Digital Secure WorkSpace Based on ' Zero Trust .' March*, 1–23. https://www.gartner.com/imagesrv/media-products/pdf/qihoo/Qihoo360-1-5SKVSAF.pdf

Gartner. (2021a). *Database Audit and Protection (DAP)*. https://www.gartner.com/en/information-technology/glossary/database-audit-and-protection-dap

Gartner. (2021b). *Network Sandboxing Reviews and Ratings*. https://www.gartner.com/reviews/market/network-sandboxing

Gartner. (2021c). *Security Information and Event Management*.

https://www.gartner.com/en/documents/3894573

Gbenga, E., Stephen, J., Chiroma, H., Olusola, A., & Emmanuel, O. (2019). Machine learning for email spam fi ltering : review , approaches and open research problems. *Heliyon*, *5*(September 2018). https://doi.org/10.1016/j.heliyon.2019.e01802

Gibert, D., Mateu, C., & Planes, J. (2020). The rise of machine learning for detection and classification of malware : Research developments , trends and challenges. *Journal of Network and Computer Applications*, *153*(January), 102526. https://doi.org/10.1016/j.jnca.2019.102526

Gordijn, B., Christen, M., & Loi, M. (2020). The Ethics of Cybersecurity. In *The International Library of Ethics, Law and Technology* (Vol. 49, Issue 0). http://www.springer.com/series/7761

Gray, C. (2020). *Top 10 AI-enabled cyber security companies*. https://aimagazine.com/technology/top-10-ai-enabled-cyber-security-companies

Gudwin, R. (1999). Reinforcement learning. *Journal of Cognitive Neuroscience*.

Hart, C. (1988). *HART Doing a literature review 1988 ch*.

Hevner, & Chatterjee. (2010). Design Research in Information Systems. In *Springer* (Vol. 28).

Hindy, H., Atkinson, R., Tachtatzis, C., Colin, J., Bayne, E., & Bellekens, X. (2020). Utilising Deep Learning Techniques for Effective Zero-Day Attack Detection. *Electronics MDPI*, 1–16. https://doi.org/10.3390/electronics9101684

Horan, C., & Saiedian, H. (2021). Cyber Crime Investigation: Landscape, Challenges, and Future Research Directions. *Journal of Cybersecurity and Privacy*, *1*(4), 580–596. https://doi.org/10.3390/jcp1040029

Hussain, H. S., Din, R., Liu, Y., Wang, Z., & Zhang, Y. (2021). Artificial Intelligence in Cyber Security. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1964/4/042072

Hussain, K. (2018). Artificial Intelligence and its Applications Goal. *International Research Journal of Engineering and Technology (IRJET)*.

ISO/IEC. (2013). *ISO/IEC 27001*.

ITU. (2008). Overview of Cybersecurity. Recommendation ITU-T X.1205. *Geneva: International Telecommunication Union (ITU)*, *1205*. http://www.itu.int/rec/T-REC-X.1205-200804-I/en

Jamal, A. A., Majid, A. M., Konev, A., Kosachenko, T., & Shelupanov, A. (2021). A review on security analysis of cyber physical systems using Machine learning. *Materials Today: Proceedings*, *xxxx*. https://doi.org/10.1016/j.matpr.2021.06.320

Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in cybersecurity. *Journal of Computer and System Sciences*, *80*(5), 973–993. https://doi.org/10.1016/j.jcss.2014.02.005

Kaloudi, N., & Jingyue, L. I. (2020). The AI-based cyber threat landscape: A survey. *ACM Computing Surveys*, *53*(1). https://doi.org/10.1145/3372823

Kamble, R., & Shah, D. (2018). Applications of Artificial Intelligence in Human Life. *International Journal of Research*, *6*(6), 178–188. https://doi.org/10.29121/granthaalayah.v6.i6.2018.1363

Kaspersky Labs. (2019). *What is Cybersecurity*. https://usa.kaspersky.com/resource-center/definitions/what-is- cyber-security

Kemmerer, R. A. (2003). Cybersecurity. *IEEE, Computer Society*, *6*.

Khan, S., & Parkinson, S. (2018). *Review into State of the Art of Vulnerability Assessment using Artificial Intelligence* (Issue September). https://doi.org/10.1007/978-3-319-92624-7

Klir, G. J., & Yuan, B. O. (1995). *Fuzzy Sets and Fuzzy Logic* (Prentice Hall (ed.)). Prentice Hall.

Lazic, L. (2019). BENEFIT FROM AI IN CYBERSECURITY. *The 11th International Conference on Business Information Security*, *October*.

Levy, Yair;Ellis, T. (2012). A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *Informing Science Journal*, *9*, 558–562. https://doi.org/10.1109/IEEM.2012.6837801

Lewis, J. A. (2006). Cybersecurity and Critical Infrastructure Protection. *Center for Strategic and International Studies*, *January*, 1–12.

Li, J. hua. (2018). Cyber security meets artificial intelligence: a survey. *Frontiers of Information Technology and Electronic Engineering*, *19*(12), 1462–1474. https://doi.org/10.1631/FITEE.1800573

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. In *Journal of clinical epidemiology* (Vol. 62, Issue 10). https://doi.org/10.1016/j.jclinepi.2009.06.006

Louati, F., & Barika, F. (2020). A deep learning - based multi - agent system for intrusion detection. *SN Applied Sciences*, *2*(4), 1–13. https://doi.org/10.1007/s42452-020-2414-z

Lu, Y. (2019). Artificial intelligence : a survey on evolution , models , applications and future trends. *Journal OfManagement Analytics*, *0012*. https://doi.org/10.1080/23270012.2019.1570365

Mikhalevich, I. F., & Ryjov, A. P. (2018). Augmented intelligence framework for protecting against cyberattacks. *Proceedings - 5th International Conference on Engineering and Telecommunication, EnT-MIPT 2018*, 143–145. https://doi.org/10.1109/EnT-MIPT.2018.00039

Ministry of Public Safety. (2010). Canada Cyber Security Strategy. *Ottawa: Public Safety Canada, Government of Canada*, 1–17. https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/cbr-scrt-strtgy/cbr-scrt-strtgy-eng.pdf

Mohamed, E. (2020). The Relation Of Artificial Intelligence With Internet Of Things : A survey The Relation Of Artificial Intelligence With Internet Of Things : A survey. *Journal of Cybersecurity and Information Management*, *March*. https://doi.org/10.5281/zenodo.3686810

Moher, D. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology*, *62*(10), 1006–1012. https://doi.org/10.1016/j.jclinepi.2009.06.005

Morgan, S. (2020). *Cybercrime To Cost The World $10.5 Trillion Annually By 2025*. Cybersecurity Ventures. https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/

Moura, J., & Trilho, P. (2021). *Not Petya Case Study* (p. 13).

Naik, B., Mehta, A., Yagnik, H., & Shah, M. (2021). The impacts of artificial intelligence techniques in augmentation of cybersecurity : a comprehensive review. *Complex & Intelligent Systems*,

0123456789. https://doi.org/10.1007/s40747-021-00494-8

Nguyen, T. T., & Reddi, V. J. (2021). Deep Reinforcement Learning for Cyber Security. *ArXiv*, *Ml*. https://doi.org/10.1109/TNNLS.2021.3121870

NIST. (2006). Minimum Security Requirements for Federal Information and Information Systems, NIST Special Publication 800-53. *Information Security*, *March 2006*(March). http://csrc.nist.gov/publications/fips/fips200/FIPS-200-final-march.pdf

NIST. (2013). : NISTIR 7298 Revision 2, National Institute of Standards and Technology. *Nist Ir*, *7298*(Revision 2), 222. http://nvlpubs.nist.gov/nistpubs/ir/2013/NIST.IR.7298r2.pdf

Oliveira, A. (2019). *Inteligência Artificial* (F. F. M. dos Santos (ed.)). Fundação Francisco Manuel dos Santos.

Ongsulee, P. (2017). Artificial Intelligence , Machine Learning and Deep Learning. *2017 Fifteenth International Conference on ICT and Knowledge Engineering*.

Oseni, A., Moustafa, N., Janicke, H., Liu, P., Tari, Z., & Vasilakos, A. (2021). Security and Privacy for Artificial Intelligence: Opportunities and Challenges. *Journal of the ACM*, *37*(4). http://arxiv.org/abs/2102.04661

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, *32*(2), 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

Oxford University Press. (2014). *Oxford Online Dictionary*. Oxford: Oxford University Press. http://www.oxforddictionaries.com/definition/english/Cybersecurity

Pachhala, N., Jothilakshmi, S., & Battula, B. P. (2021). A Comprehensive Survey on Identification of Malware Types and Malware Classification Using Machine Learning Techniques. *Proceedings of the Second International Conference on Smart Electronics and Communication*, 1207–1214.

Pedamkar, P. (2020). *What is AI*. EDUCBA. https://www.educba.com/what-is-artificial-intelligence/?source=leftnav

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Peslak, Alan; Hunsinger, S. (2019). What Is Cybersecurity and What Cybersecurity Skills Are Employers Seeking? *Issues In Information Systems*, *20*(2), 62–72. https://doi.org/10.48009/2_iis_2019_62-72

Razzaq, A., Hur, A., Ahmad, H. F., & Masood, M. (2013). Cyber Security : Threats , Reasons , Challenges , Methodologies and State of the Art Solutions for Industrial Applications. *IEEE Eleventh International Symposium on Autonomous Decentralized Systems*.

Read, M., Andrews, P., & Timmis, J. (2012). An Introduction to Artificial Immune Systems. *Research Gate*, *June 2015*. https://doi.org/10.1007/978-3-540-92910-9

Rosenberg, I., Shabtai, A., Elovici, Y., & Rokach, L. (2021). Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. *ACM Computing Surveys*, *54*(5). https://doi.org/10.1145/3453158

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: a Modern Approach* (L. Pearson Education (ed.);

4th ed.). Pearson Education, Limited.

Sagar, B. S. (2019). Providing Cyber Security using Artificial Intelligence – A survey. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, *Iccmc*, 717–720.

Salim, S. (2019). *Revealed: The biggest data breaches of 2018*. Digital Information World. https://www.digitalinformationworld.com/2018/12/biggest-data-breaches-of-2018.html

Saltzer, H., & Schroeder, M. (1975). The protection of information in Computer Systems. *IEEE, Computer Society*, *63*(Iccas), 1233–1237.

Sarker, I. H., Md, ·, Furhad, H., & Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. *SN Computer Science*, *2*, 173. https://doi.org/10.1007/s42979-021-00557-0

Scott, J. (2017). Signature Based Malware Detection is Dead. *Cybersecurity Think Tank, Institute for Critical Infrastructure Technology*, *February*.

Şeker, E. (2019). Use of Artificial Intelligence Techniques / Applications in Cyber Defense. In *arXiv*.

Shaukat, K., Luo, S., Varadharajan, V., & Hameed, I. A. (2020). Performance Comparison and Current Challenges of Using Machine Learning Techniques in Cybersecurity. *Energies MDPI*.

Shaukat, K., Luo, S., Varadharajan, V., & Member, S. (2020). A Survey on Machine Learning Techniques for Cyber Security in the Last Decade. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.3041951

Shenfield, A., Day, D., & Ayesh, A. (2018). Intelligent intrusion detection systems using artificial neural networks. *ICT Express*, *4*(2), 95–99. https://doi.org/10.1016/j.icte.2018.04.003

Shirey, R. (2007). *Request for Comments 4949*. Network Working Group. https://www.rfc-editor.org/rfc/rfc4949.html

Simon, H. A. (1996). The Sciences of the Artificial. In *Technology and Culture* (Vol. 11, Issue 1). https://doi.org/10.2307/3102825

Smith, R. E. (2012). A Contemporary Look at Saltzer and Schroeder ' s 1975 Design Principles. *IEEE, Computer Society*, *December*, 20–25.

Sosin, A. (2018). How To Increase the Information Assurance in the Information Age. *Journal of Defense Resources Management*, *9*(1), 45–57. http://capella.summon.serialssolutions.com.library.capella.edu/2.0.0/link/0/eLvHCXMwrV1LS8QwEB58IAgiiopv8gdW85h005tVutaDLNiKx9Jk0uMisv7_TZqCr4MHPSaBGSYhmZnkyzcASl7xybczwZnOCY8kCf3UoemJpDVoXHB3ytvs6wPvR-WvCBFLbMFpHq8lWUvcWTVFRMd7qwUp6jPSTqLr9TpshoglxRcxcp

Strain, L. (2018). *The seven most colossal data breaches of 2017*. Malware Bytes Lab. https://blog.malwarebytes.com/cybercrime/2017/12/the-seven-most-colossal-data-breaches-of-2017/

Suliman, S. I., Safwan, M., Shukor, A., Kassim, M., & Mohamad, R. (2018). Network Intrusion Detection System Using Artificial Immune System (AIS). *2018 3rd International Conference on Computer and Communication Systems Network*, 178–182.

Tahir, R. (2018). A Study on Malware and Malware Detection Techniques. *International Journal of*

Education and Management Engineering, *8*(2), 20–30. https://doi.org/10.5815/ijeme.2018.02.03

Thrall, J. H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., & Brink, J. (2018). Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success. *Journal of the American College of Radiology*, *15*(3), 504–508. https://doi.org/10.1016/j.jacr.2017.12.026

Truong, T. C., Diep, Q. B., & Zelinka, I. (2020). Artificial Intelligence in the Cyber Domain: Offense and Defense. *MDPI /Symmetry*. https://doi.org/10.3390/sym12030410

Upadhyay, I. (2020). *Top 10 Challenges of Cyber Security Faced in 2021*. https://www.jigsawacademy.com/blogs/cyber-security/challenges-of-cyber-security/#Software-Vulnerabilities

Uttar, I., Section, P., Conference, I., & Engineering, C. (2018). A Survey Paper on Cryptography Techniques. *International Journal of Computer Science and Mobile Computing*.

Vaishya, R., Javaid, M., Khan, I., & Halem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. *Elsevier*.

Valente, A., & Murphy, R. (2020). Governance, Risk, And Compliance Platforms. In *The Forrester Wave*.

Veiga, A. P. (2018). *Applications of Artificial Intelligence (AI) to Network Security* (Issue March).

Verma, M. (2018). Artificial intelligence and its scope in different areas with special reference to the field of education. *International Journal of Advanced Educational Research*, *3*, 2455–6157. www.educationjournal.org

Vieira, J., Dias, F. M., & Mota, A. (2004). Neuro-Fuzzy Systems : A Survey. *5th WSEAS NNA International Conference on Neural Networks and Applications, Udine, Italia*.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii. https://doi.org/10.1.1.104.6570

Wiafe, I., Koranteng, F. N. T. I., Obeng, E. N., Wiafe, A., Gulliver, S. R., & Assyne, N. (2020). Artificial Intelligence for Cybersecurity : A Systematic Mapping of Literature. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.3013145

Yadav, V. (2020). A Study of Threats , Detection and Prevention in Cybersecurity. *International Research Journal of Engineering and Technology (IRJET)*, *May*, 1150–1153.

Yamin, M. M., Ullah, M., Ullah, H., & Katt, B. (2021). Weaponized AI for cyber attacks. *Journal of Information Security and Applications*, *57*(January), 102722. https://doi.org/10.1016/j.jisa.2020.102722

Yeasmin, S. (2018). Benefits of AI in Medicice. *2nd International Conference on Computer Applications & Information Security, ICCAIS'2019*.

Zeadally, S. (2020). Harnessing Artificial Intelligence Capabilities to Improve Cybersecurity. *IEEE Access*.

Zelonis, J., & Lyness, T. (2019). Vulnerability Risk Management. In *The Forrester Wave*.

Zhang, Y., Li, P., & Wang, X. (2019). Intrusion Detection for IoT Based on Improved Genetic Algorithm

and Deep Belief Network. *IEEE Access*, *7*, 31711–31722. https://doi.org/10.1109/ACCESS.2019.2903723

Zhao, J., Masood, R., & Seneviratne, S. (2002). A Review of Computer Vision Methods in Network Security. *IEEE Communications*, *23*(3), 1–34.

ANNEXES

Interviews conducted with experts in this area

Interview 1: *Julian Williams*, Global Security Designer @ Vodafone Group, date: 15.03.2022

Question 1: Do you consider this framework is useful and why? If not, could you explain the reasons?

So, the framework is useful from an overall high-level concept of the implementation of AI technology. It could be used as an overall starting point for the discussion of implementation of AI technology.

Question 2: Do you have any criticism towards the proposed framework? Could you provide an explanation?

The framework is very generic, it does not focus on a single technology, but this is what the high-level framework is for then. That is my only critique. My only criticism is that the AI technology capability across security is immense, you could use it anywhere. So, the framework covers really anything, yeah…

Question 3: Would you consider implementing the proposed model? Could you explain why or why not?

If the business was progressing to implement or use AI technology, then the framework would be good to provide a scope of where in the business to implement the AI technology and what threats and risk reduction would be achieved. So, the framework and all the investigation around this, the controls against you have the reference says you have the risk reduction etc. So, the framework would provide, yes, an overview of where the AI technology would be consumed and, also the risks… We could do risk reductions, yeah…

Question 4: Do you have any recommendations for further improvements of this framework?

I've done a couple, well, we would… We would put the encryption technologies, removal of some threats that don't pertain to the risk of access control and IA, like fishing and malware, and adding password attacks in here. Conditional access policies are a vendor specific branding, so you could probably use terminology related to zero trust model here. Sandbox should probably be removed from the network technologies as this is a capability, not specifically a network technology, as if an AI solution detects the threat it puts it in a sandbox for analysis, the AI here actions the activity. The diagram described as solution architecture would be better referenced as something like "stages of an AI Architecture", I wouldn't consider that as a security architecture diagram. But overall, the framework is very good. Good work!

Interview 2: *Marco Reis*, Cybersecurity Professor @Nova IMS, date: 23.03.2022

Question 1: Do you consider this framework is useful and why? If not, could you explain the reasons?

It's useful indeed but would be important to understand how it could be used for example for detection of the percentage of critical incidents and reduction of the false positives rate, because these are metrics which in terms of SOC have implications at the productivity and risk management levels. Regarding costs when we talk about artificial intelligence in cybersecurity normally this means an increase and not a reduction of costs, because when organization opt for an AI based solution often means subscribing to a premium service. Organizations normally look at this as a mechanism to improve productivity: Reduction in the time spent on detection and response to security incidents.

Question 2: Do you have any criticism towards the proposed framework? Could you provide an explanation?

Criticism towards the framework… I only looked at the framework in a broad way, for now I don't think there is much criticism to point out. The main question here is that there are a wide set of AI techniques which can be used, and perhaps they are too many… But a framework is this, it's something broad and given the complexity and coverage of the subject it's difficult to have a more specific scope. Besides, the utilization of artificial intelligence in this area it's very focused on machine learning and information regarding the utilization of other techniques are normally more addressed in academic terms. We're far from being able to utilize other AI techniques in the cyber security area that are not machine learning.

Question 3: Would you consider implementing the proposed model? Could you explain why or why not?

When we use a framework, normally we must have an objective from the point of view of security in mind. We must consider that we've got a set of assets which we must protect and for that it's required a certain investment, and at this moment is very difficult to say if this framework implementation would be effective to reduce the risk or not. But it's very likely that it would because AI mechanisms are more effective. But if organizations need to pay for a premium service to get an AI based solution, then this is a problem for the decision makers of the organizations, because the security budget must be limited to the value of the assets to protect. On this point would also be interesting, and this probably is more related with the next question, understand… It's very important to understand to which extent it makes sense or not to pay for a premium service, that is, understand that if the cost of the premium is too high, the assets have enough value that justifies protecting these assets using an AI based solution.

Question 4: Do you have any recommendations for further improvements of this framework?

In terms of suggestions for improvement, what I can suggest is perhaps the budget topic being addressed in the beginning… But I'm also not sure if it's possible to define a budget before knowing what will be necessary… Not such an initial budget definition but possibly try to understand what would be the top limit that can be used as a budget. Although sometimes this can't be done exactly like this because there is the risk that we end up with a very high budget but having in consideration what are the risks of a certain incident to occur and if that incident results in a high business impact, understand how we could start from the risk component analysis towards what would be the security architecture. This would be my suggestion for the initial methodology… This isn't far from the methodology presented here, but I would say that risk evaluation versus the decision of investing in a premium service should be done in the strategy phase. There is a document from Gartner, if I'm not

mistaken, which recommends organizations in the initial phase not to use AI solutions just because they're AI solutions, but only if it makes sense to use them considering the threats which they may be subject to. In general, I think this work is well done, and a lot of information about the topic has been collected. This is an area where there is nothing yet, but we think that in 2 to 10 years' time frame more AI techniques start being used. Nowadays, this is very used in terms of marketing by the vendors, which say that they've got AI based solutions, but when people ask: So, what does that do? Often the vendors are not capable of answering this in an objective way…

Interview 3: *Nuno Pires*, Cybersecurity Consultant & Trainer @Centro Nacional de Cibersegurança, date: 19.04.2022

Question 1: Do you consider this framework is useful and why? If not, could you explain the reasons?

Yes, this Framework is useful. Why? Because it can be a complementary implementation in relation to the Portuguese National Cybersecurity Framework defined by Portuguese National Cybersecurity Centre. The Portuguese National Cybersecurity Framework summed up a set of well-known Cybersecurity international accepted standards which allows organizations to perform a risk-based approach to tackle cyber threats, establishing the foundations for the implementation of security measures for networks and information systems. For that purpose, specific measures have been identified and structured around the following five Cybersecurity objectives: identification, protection, detection, response and recover phases for handling cybersecurity incidents, including the necessary organizational environment to cope with them. With the implementation of the AI based techniques of the proposed framework, the response using threat intelligence will step up to the next level of security.

Question 2: Do you have any criticism towards the proposed framework? Could you provide an explanation?

This proposed framework shows the macro approach to apply AI to cybersecurity. It would be interesting, in further studies, to view subset of AI like ML integrated and detailed in the framework.

Question 3: Would you consider implementing the proposed model? Could you explain why or why not?

With the necessary adaptation to Portuguese National Cybersecurity Framework, and some further development to perform a case study, it could be possible to apply it in most of the public government infrastructure.

Question 4: Do you have any recommendations for further improvements of this framework?

For proof of concept, and from with my knowledge and experience in this subject, it covers all the major aspects needed to apply AI to cybersecurity.

Interview 4: *Ana Filipa Monteiro*, Cybersecurity Information Consultant @ ORAMIX, date: 22.04.2022

Question 1: Do you consider this framework is useful and why? If not, could you explain the reasons?

Yes, I find it useful.

Question 2: Do you have any criticism towards the proposed framework? Could you provide an explanation?

I don't have anything special to criticize Because it is well designed, and it covers all the technologies and risks associated with them.

Question 3: Would you consider implementing the proposed model? Could you explain why or why not?

Yes, this framework would help the operations to be smother and less stressful on the team.

Question 4: Do you have any recommendations for further improvements of this framework?

In my opinion, there are 2 things missing: On the Implementation phase it's missing the initial budget and on the roadmap phase, it's missing the problem/solution to report to the stakeholders to define new budget.