Artificially intelligent cyberattacks

Erik Zouave, Marc Bruce, Kajsa Colde, Margarita Jaitner, Ioana Rodhe, Tommy Gustafsson

FOI-R--4947--SE

March 2020

FOI

Erik Zouave, Marc Bruce, Kajsa Colde, Margarita Jaitner, Ioana Rodhe, Tommy Gustafsson

Artificially intelligent cyberattacks

Titel	Artificially intelligent cyberattacks
Title	Artificially intelligent cyberattacks
Rapportnr/Report no	FOI-R--4947—SE
Månad/Month	March
Utgivningsår/Year	2020
Antal sidor/Pages	48
ISSN	1650-1942
Kund/Customer	Totalförsvarets forskningsinstitut FOI
Forskningsområde	Informationssäkerhet
FoT-område	Inget FoT-område
Projektnr/Project no	I149181
Godkänd av/Approved by	Lars Höstbeck
Ansvarig avdelning	Försvarsanalys

Bild/Cover: Marc Bruce

Summary

This report explores the possibilities and applications of artificial intelligence (AI) within the various stages of a cyberattack. It is a literature review of the current state of the art in AI application within the cyber domain, with specific focus on research from the antagonist perspective and technology that can be used both defensively and offensively. The five stages referred to in this report include reconnaissance, access and penetration, internal reconnaissance and lateral movements, command, control and exfiltration and sanitation. Within the report, 19 use cases for AI were found. The antagonistic use cases demonstrate AI versus human technology user and AI versus technology scenarios, with the latter including dimensions of an AI-supported attacker versus AI-supported defender arms race. The found use cases demonstrate that there is a higher technology readiness level at early phases of the cyberattack. Moreover, the sources reviewed imply that some AI-supported cyberattacks can already be observed in the wild. The report also identifies four strategic capabilities that antagonists may achieve through the malicious use of AI in cyberattacks; namely aggregation, repetition, deception, and manipulation.

Keywords: Artificial intelligence, cyberattacks, anatomy, automation, offensive cyber, malicous artificial intelligence

Sammanfattning

Denna rapport utforskar möjligheterna av att utnyttja artificiell intelligens (AI) inom ett dataangrepps olika faser. Det är en avskanning av forskningsläget gällande AI och dess applikationer på cyberdomänen, med särskild fokus på forskning från det antagonistiska perspektivet och forskning på teknologi som kan användas både defensivt och offensivt. De fem stadier som rapporten sorterar användningsfallen utifrån omfattar rekognosering, intern rekognosering och lateral spridning, fjärrstyrning och exekvering mot mål samt exfiltration och sanering. Inom ramen för studien identifierades 19 användningsfall för AI i anatomin av ett dataangrepp. Fallen påvisar scenarion av AI mot mänsklig teknik-användare och AI mot teknik där den senare även omfattar aspekter av en kapprustning mellan AI-stödd angripare och AI-stödd försvarare. Dessa användningsområden påvisar en högre mognadsgrad i tidigare skeden av anatomin. Dessutom antyder källorna att AI-stödda dataangrepp redan har observerats. Ytterligare identifierar studien fyra strategiska förmågor som hotaktörer kan uppnå med AI-stödda dataangrepp; aggregering, repetition, falsifiering, och manipulation.

Nyckelord: Artificiell intelligens, dataangrepp, automatisering, anatomi, offensiv cyber, skadlig artificiell intelligens

Contents

1 Introduction

Artificial intelligence (AI) will play an increasingly important role in the execution of future cyberattacks. The works of numerous security researchers indicate this prediction (Guarino, 2013; UNIDIR 2017a and 2017b; Brundage et al, 2018; Horowitz et al, 2018). It has further been hinted at by research and development projects launched within the IT-industry (Kirat et al, 2018) and defense industrial complexes (DARPA, 2017). The 2016 DARPA Grand Cyber Challenge, pitting machine against machine in the detection, exploitation, and mitigation of system vulnerabilities marked an important milestone for artificially intelligent cyberattacks (DARPA, 2015 and 2016). While events such as these may sound like fantastic and recent feats in technological development, Dheap (2017) traces the application of AI in cyber security, particularly naïve Bayes classifiers, back to the 1990's efforts to develop spam-filtering technology. Moreover, recent leveraging of machine learning by malicious actors in actual cyberattacks to mimic normal network behaviors (Norton, 2017; Darktrace, 2018) and in applied experiments to automatically discover, profile, evaluate and engage (spear phish) targets on social media (Seymour & Tully, 2016) presents a stark proof of concept for malicious use. Researchers are currently approaching this topic from disparate disciplines, angles, and motivations ranging from questions of how AI might be implemented in future cyberattacks (Guarino, 2013; Dheap, 2017), how to implement AI in specific attack-related use cases (Juric et al, 2019; Löfvenberg et al, 2019; Greeff & Ross, 2019; Grant, 2018; Falco et al, 2018; Cakir & Dogdu, 2018; Wirkuttis & Klein, 2017), how AI has been implemented in attacks (Darktace, 2018), to what the international legal policy concerns are (UNIDIR, 2017; Brundage et al, 2018; Scharre and Horowitz, 2018). Notably, recent research by the NATO Hybrid Centre of Excellence (2019a and 2019 b) highlights the role of AI in hybrid warfare. A central proposition of this report is that cyberattacks can be separated into various stages, constituting the cyberattack anatomy (Pfleeger 2010; Oracle 2017), and that AI as a set of technologies has a role to play in that anatomy (Guarino, 2013).

This report summarizes the results of a review of research literature on AI in cyberattacks. The Swedish Defence Research Agency conducts literature reviews on a yearly basis indirectly financed by the Swedish Armed Forces. The reviews explore international and Swedish research as well as the implications of the research for Swedish defense and the research conducted within the Swedish Defence Research Agency itself. Within the context of Swedish information and cyber security, research should contribute to effective, robust, and sustainable approaches to digital security and integrity (Swedish Government Offices, 2016). This report is directed at the Swedish competent authorities in the domain of cyber security and their international partners, in particular persons responsible for the development of technical expertise in these organizations. The report is further directed at the AI research community.

1.1 Purpose

This report explores the possibilities and applications of artificial intelligence within the anatomy of cyberattacks. It asks four main research questions:

1. *How can AI be implemented through the stages of the cyberattack anatomy?*
2. *Which are the predominant AI technologies that are necessary for these implementations?*
3. *What is the technology readiness (i.e. idea, prototype, validation, product) of various forms of AI augmentation in the cyberattack anatomy?*
4. *What are the restrictions and obstacles to the implementation of AI in cyberattacks?*

1.2 Summary of findings

The literature review in this report reviews 96 sources that either takes the antagonist's perspective or concerns the application of dual use technology. By dual use technology, we mean technology that can be used for both defensive and offensive purposes. Within this literature, the study identified 19 use cases for AI in the anatomy of cyberattacks. The development of certain AI technologies and the discrepancies in experimental research indicates that some use cases for AI within the anatomy of cyberattacks are currently more mature than others. This is demonstrated by the existence of commercially available systems (typically security implantations of dual use technology), the identification of real AI-supported attacks, experimental research on prototype solutions, and conceptual research. Use cases with higher technology readiness levels are generally exhibited in applications for the earlier stages of the anatomy, especially reconnaissance, as shown in figure 1. This figure demonstrates an estimated mean technology readiness across use cases within phases in the cyberattack anatomy. The simplified version of the technology readiness scale is explained in section 4.3 of this report. The use cases presented in the results of this report should be regarded as a snapshot of currently known areas of research, rather than a prediction for future malicious end use. However, sources to this report identified AI-supported cyberattacks in the wild, where AI was used in profiling, phishing, captcha defeats, and network analysis.

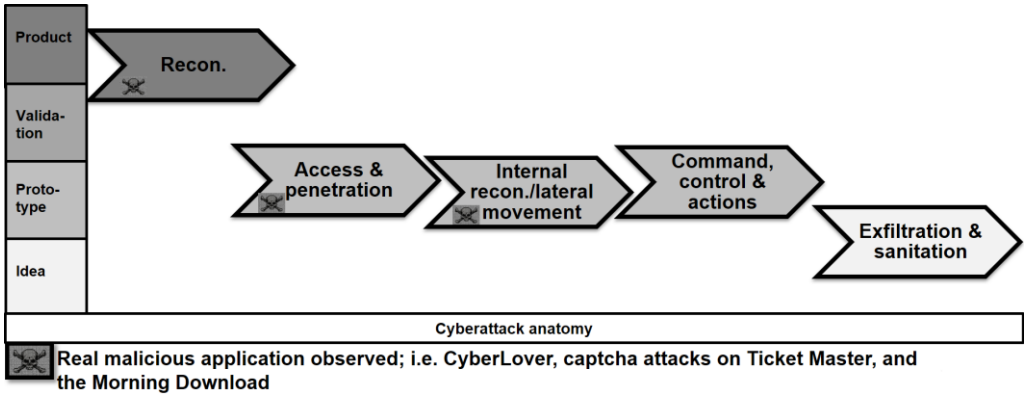

Figure 1 Overall technology readiness level for AI use in the cyberattack anatomy.[1]

1.3 Disposition

The remainder of this report presents three selected cases of AI-supported cyberattacks (section 2), methods and limitations to the study (section 3), background information on AI technologies and the anatomy of cyberattacks (section 4), as well as previous AI research at the Swedish Defence Research Agency (FOI) (section 5). The report then describes the results of the literature review, notably on the application of AI within reconnaissance, access and penetration, internal reconnaissance and lateral movement, command and control and payload delivery, and exfiltration and sanitation (section 6). Finally, the findings are discussed (section 7) and the conclusions and suggestions for future research are presented (section 8).

[1] Skull from Shutterstock: https://www.shutterstock.com/image-vector/skull-crossed-bones-danger-piracy-sign-550387903?irgwc=1&utm_medium=Affiliate&utm_campaign=Pixabay+GmbH&utm_source=44814&utm_term=https%3A%2F%2Fpixabay.com%2Fsv%2Fimages%2Fsearch%2Fskull%2520computer%2F

2 Three cases: AI-supported cyberattacks

In this section, we introduce three example cases where AI has or could potentially be leveraged to perform tasks previously done by human attackers. The examples are meant to be illustrative, and to provide context to research results presented later in this report. By describing a few examples of how the technology has been used or its use is imagined, it is easier to see the use cases as more than just lists of research areas of interest to scholars. The examples should not be regarded as comprehensive but have rather been selected as having "most different" qualities (from each other), and therefore being demonstrations of potential AI-capabilities. Other cases, exist and have been known to researchers for some time. For example, "CyberLover" (Rossi, 2007) was a natural language processing bot that profiled and attempted to trick visitors in dating chat rooms to click on malicious links. Further, machine learning was allegedly used in a case of wire fraud and hacking, defeating captchas on the Ticketmaster website to automatically purchase over 1 million tickets for resale (Bursztein et al 2011; McMillan, 2010; Zetter, 2010). The Morning Download, is likewise said to have employed machine learning to detect and mimic patterns on its target network, thereby avoiding and hiding from countermeasures (Norton, 2017). The three selected example cases range from an AI-supported spear phishing, autonomous software systems designed to seek out software vulnerabilities and exploit them, and next-generation cryptoworms.

2.1 SNAP_R

While many internet users are aware that they should avoid clicking on unknown links in emails or open unknown attachments, this advice is less seldom followed on social media because the medium tends to exhibit higher degrees of trust between users (Waddell, 2016). In 2016, two cyber security researchers from ZeroFox released a proof-of-concept program designed to show how social media users can be exploited using AI. This tool, named Social Network Automated Phishing and Reconnaissance (SNAP_R), works in two stages; target discovery and automated spear phishing. During the target discovery phase, users are categorized into clusters based on publicly available data such as their profile, follower interactions, and engagement metrics (Seymour & Tully, 2016). In the second phase, automated spear phishing, the program analyses tweets by twitter users belonging to the cluster it believes is most vulnerable to a social engineering attack and determines what they post about. It then automatically creates highly relevant replies using two forms of deep learning models; long short-term memory (LTSM) and Markov chains (Seymour & Tully, 2016). The bot can attach a shortened obfuscated link within its replies where it can attempt to 'phish' the target. Successful phishing attempts occur when a target clicks on the link and enters in personal details, such as a username and password, into what ostensibly appears to be a legitimate website but is in reality an attacker's own reproduction. With the user credentials captured, the attacker can then attempt to access and potentially take over the target's account. SNAP_R generated attacks have been shown in tests to have a click-through rate of over 30%, which is significantly higher than the 5-14% previously reported for large-scale phishing campaigns (Seymour & Tully, 2016).

2.2 2016 DARPA Cyber Grand Challenge

The 2016 Cyber Grand Challenge (CGC) featured advanced software systems known as cyber reasoning systems (CRS). The challenge was to see if they could compete against each other in detecting, analyzing, exploiting, and repairing software vulnerabilities autonomously and in real-time. The event, funded by the American Defense Advanced Research Projects Agency (DARPA), was marketed as "the first-ever tournament for fully automatic network defense systems" (DARPA, 2015) with the winning team being awarded

$2 million USD. DARPA's desire to hold such a tournament was born out of the need to, inter alia, detect software vulnerabilities faster than presently possible through manual labor (DARPA, 2015). By implementing automated vulnerability detection and patching systems it is hoped that this time lag can be reduced to mere milliseconds, thus mitigating the damage potential of future cyberattacks (Avgerinos, et al., 2018). While indeed a significant undertaking, the CGC made inroads towards this goal when the seven finalist teams demonstrated that it was indeed possible for CRS to find, patch, and even autonomously attack competitor systems with software vulnerabilities (BBC News, 2016). Participating teams included researchers from academia, e.g. University of California and University of Ohio, as well as industry contestants, such as Raytheon and GrammaTech Inc. The winning team, which found and patched the most vulnerabilities while also using them to hamper the efforts of the other competitors was Pittsburgh-based cybersecurity developer ForAllSecure with their CRS 'Mayhem' (DARPA, 2016).

2.3 Intelligent cryptoworms

Researchers such as Stoecklin (2018) and Lando (2018) envision a possible future of AI-supported cryptoworms. A worm is a malicious code that possesses the ability to propagate from one system to another by itself, with no human intervention. It is both difficult to contain as well as eradicate the infection. A cryptoworm both self-propagates and encrypts files as one of the command and control functions (Svensson et al, 2019). Famous cryptoworms, WannaCry and NotPetya, encrypted computer files, affecting systems around the world (Svensson et al, 2019; Fruhlinger, 2017). The intelligent cryptoworm scenario is more hypothetical than SNAP_R and the DARPA CGC examples. Apart from projects at IBM (Kirat et al, 2018), a major corporation in the field of information technology, the scenario currently lacks support in experimental research. Kirat et al (2018) visualise AI, specifically deep neural networks (DNN) being used in cryptoworm target selection, based on target attributes such as geolocation, software and user activity. With intelligent targeting, a malware can impersonate human interactions, i.e. intelligent phishing emails (Lando, 2018). Rhodes (2019) and Kirat et al (2018) imagines either machine learning (ML) or deep neural networks (DNN) used to conceal a ransomware attack, and possibly to manipulate the target system into disabling security measures. Hence, with automated self-propagation, Rhodes (2019) foresees that it would be possible to infect a large number of systems while slowing down or avoiding detection by changing the identifiable features of the malware (Rhode, 2019). According to Stoecklin (2018), the malware might be able to hide itself inside other applications to avoid detection. These evasive characteristics would make it challenging for defenders to identify patterns of a threat (Rhode, 2019). Moreover, a cryptoworm that is supported by AI would produce attacks faster according to Dixon and Eagan (2019) or make the attacks more efficient, according to Batt (2019).

3 Methods, limitations, and challenges

This study predominantly relied on a review of existing research literature in English (and some in Swedish) to identify use cases for AI in the stages of cyberattacks. This included academic, peer-reviewed literature, research reports by governmental and non-governmental organizations, experimental and prototype reporting from researchers and industry, as well as investigative and other reporting from IT-security firms. The review was processed through four major steps.

The first step was the initial collection of literature from Google, Google Scholar, Scopus, containing the following search terms:

> "Cyber operation" OR "hacking" OR "computer network exploitation" OR "cyberattack" OR "cyber attack"
>
> AND
> "reconnaissance" OR "research" or "access" OR "infilt*" "exploit*" OR "penetrat*" OR "lateral" OR "expan*" OR "command" OR "control" OR "exfilt*" OR "sanit*"
>
> AND
> "AI" OR "artificial intelligence" OR "neural" OR "Bayes*" OR "learning" OR "heuristic" OR "support vector" OR "swarm" OR "evolution*" OR "Markov decision process" OR "generative adversarial" OR "fuzzing" OR "natural language processing".

The second step was an initial assessment and structuring of the identified literature. The assessment structured publications according to their relevance to the main research questions of this report. The results of the collection and structuring are provided in table 1.

Table 1 Initial literature collection results

Search engine	Relevant sources	Irrelevant sources	Total
Google	5	3	8
Google Scholar	11	39	50
Scopus	10	142	152
All	26	184	210

The third step was a workshop on tentative results from the initial collection. The workshop gathered an interdisciplinary group of researchers from FOI to discuss the overall findings from the initial collection and assessment and the way forward with the study. The workshop identified several hypothetical use cases for AI-supported stages in the cyberattack anatomy (table 2). These hypothetical use cases were then followed up with yet another round of data collection from databases.

Table 2 Hypothetical use cases for AI in the anatomy of cyberattacks

Step in the anatomy	Hypothetical uses-cases for AI
Reconnaissance	▪ Intelligence collection ▪ User profiling ▪ Vulnerability detection
Access and penetration	▪ Attack planning ▪ Social engineering ▪ Attack code generation
Internal reconnaissance and lateral movement	▪ Network behavior analysis
Command and control	▪ Data exfiltration
Exfiltration and sanitation	▪ Evidence erasure

During the fourth step, additional academic literature was collected from Scopus, Google Scholar and IEEE on the specific topics of the identified hypothetical use cases. The total number of relevant publications collected was 96. These publications were collected from the reference lists of sources and from the renewed collection through the search engines. Renewed searches were conducted for identified use case separately and new searches were conducted each time a new use case emerged from the literature. The literature was selected for this report when it was either being conducted from the antagonist perspective, from an offensive operations perspective, or from a defensive perspective but with technology that can be used for both defensive and offensive operations. The initial literature and the expanded literature was then processed to identify the 19 final use cases for AI. The use cases were identified through instances in the literature describing experiments or ideas about how various AI techniques can be used to affect technology users, systems or networks in a way that is negative for their security. The identification of use cases contributed to answering the first question of how AI can be implemented in the stages of a cyberattack. Furthermore, information about the AI technologies used were collected to answer the second research question regarding predominant AI technologies.

Finally, the identified use cases were also correlated to prototype experimentation in the research literature as well as existing security industry solutions, actual cases of malicious development, and end use through cyber security industry websites, security blogs and news sources interviewing or citing security practitioners.

The literature review method is associated with several challenges. The overarching challenge for a study of this scope is to adjust the research method to a relevant level of detail. On one hand, some of the found use cases for AI could merit a concise and restricted literature review of their own with specific methodological adjustments (e.g. Löfvenberg, Sommestad & Bildsten, 2019). On the other hand, some AI use cases may have more fundamental research, more thoroughly developed research, or more useful solutions outside the specific study of cyberattacks or cyber security. Another key challenge for the study is formulating a method that can account for conceptual inconsistencies in the literature, given both the lack of a universal definition of AI and the variety of the literary sources. A notable problem for a review of scientific sources over varied topics is accounting for the multitude of categories and sub-categories of potential AI-techniques used by researchers. For this reason, an interdisciplinary group of researchers formulated search terms together, based on previous research from different fields of study. An additional challenge is the potential restrictiveness of a potentially replicable literature collection. While it is possible to use literature collected with predetermined search terms from predetermined databases only, there is no guarantee that the results will yield the most relevant and authoritative works. Hence, once hypothetical use cases were identified, the study no longer restricted search term and database use. A related challenge to the replicability of the literature collection method is that Google adjusts search results to individual user preferences, providing different suggestions to different researchers.

4 Terminology: AI in the anatomy of cyberattacks

Researchers and IT-security companies have previously described the process underpinning cyberattacks in terms of malicious activities in various phases of the attack (Pfleeger 2010; Lockheed Martin 2015; MITRE, 2015; Oracle 2017). In 2013, Guarino's (2013) research alluded to the inclusion of AI within various steps of that process. Prior to presenting the results of the literature review in this report, it is necessary to briefly explain what is meant by AI in this report, the steps in the cyberattack anatomy, and the understanding of technology readiness that frames the results of the literature review.

4.1 Artificial intelligence (AI)

Traditional conceptualizations of AI are closely tied to emulating human reasoning, or even cognition (Bellman, 1978; Charmiak & McDermott 1985; Haugeland, 1985). The focus on human reasoning persists to some degree today, with the pursuit of *artificial general intelligence* (AGI) (Goertzel & Pennachin, 2007); an intelligence as versatile and capable as the human mind. However, near-human cognitive performance is not a useful threshold for this report.

An alternate way to define "intelligence" in machines is that it has specific goal-oriented and problem-solving (*narrow*) applications (McCarthy, 20017). Studying the narrow field of AI application a core paradigm for the technology includes the automated capacity to *sense*, *think* and *act*, (Siegel, 2003; Teahan, 2010) or alternately *observe*, *orient*, *decide*, and *act* (Boyd, 1987; Beran et al, 2017). These paradigms are implemented through various technologies collectively referred to as AI.

This report takes a methodological approach to classifying AI, previously used in horizon scanning conducted at the Swedish Defence Research Agency (Schubert, 2017; Svenmarck et al, 2020). This means that we base our conclusions on sources citing specific AI technologies and methods, including, but not limited to, the following:

Machine learning (ML) or the use of statistical algorithms to detect patterns and relations in data, and make predictions, based on prior training on datasets. Machine learning models include techniques such as Bayesian reasoning and regression analysis, classifiers such as support vector machines (SVM), and predictive modelling, such as decision trees (Murphy, 2010). Machine learning has also been combined with other automation techniques in computer security research, e.g. neural fuzzing which attempts to improve automated vulnerability detection (Blum, 2017).

Artificial neural network (ANN) and deep learning (DL) are a form of machine learning with applications such as object classification, detection, and machine control (Goodfellow, Bengio & Courville, 2016). Generative adversarial networks (GAN) is a neural network technology that has recently become associated with deep fakes and fraudulent data replication (Choi et al, 2018). With GAN, two neural networks, a generative network and a discriminative network, are used to recreate features in content, evaluate those features and over trials improve the realism of how those features are represented by the machine.

Natural language processing (NLP) involves machine learning, statistical models and other techniques as well as linguistics to analyze human language (Mitkov, 2003; Indurkhya & Damerau, 2010). It is a predominant technology in processing and analysis of, inter alia, text-based digital data.

Swarm intelligence (SI) involves the machine emulation of flock behaviors (Ahmed & Glasgow, 2012). Swarm technology is thus associated with the coordination, formation and organization of intelligent agents and computerized systems (Svenmarck et al, 2020).

It should be noted that the literature reviewed in this report contains a greater number and variety of AI technologies than can reasonably be explained in detail here.

4.2 Cyberattack anatomy

By cyberattack, we are referring to the antagonistic use of malicious code, programming or other information technology tools to negatively affect the security of technology and its users. A central proposition of this report is that cyberattacks can be separated into various stages and that AI as a set of technologies has a role to play in that anatomy (Guarino, 2013). Researchers and the IT-security industry have posed several similar, but not identical, frameworks and terminologies for the separation of the various phases, parts and components of cyberattacks. These frameworks include the cyber kill chain (Lockheed Martin 2015), Mitre's (2015) ATT&CK framework, and the idea of a cyberattack anatomy (Pfleeger 2010; Oracle 2017). For the purpose of structuring the results and findings of this report, the term "anatomy" will be used. This anatomy encompasses five steps adapted specifically for this report; *reconnaissance*, *access and penetration*, *internal reconnaissance and lateral movement, command and control,* as well as *exfiltration and sanitation* (Pfleeger, 2010; Oracle, 2017; Advanced Networks Systems, 2018; Metivier, 2018).

Reconnaissance is the process of research that a malicious actor undertakes to collect intelligence on potential targets, including the strategic evaluation of potential targets, or a particular designated target. This step may involve profiling the potential target or targets. The profile might include target behaviors, their strengths and weaknesses in ensuring the security of their systems, the persons involved in the social network of the target, as well as relevant technical information about the systems themselves such as IP-addresses.

Access is the means by which a malicious actor leverages knowledge about the target to select and apply an appropriate propagation method to penetrate the target. A potential means of access is the use of e-mails with infected attachments or providing fraudulent links.

Internal reconnaissance and lateral movement is the development of the initial reconnaissance and access steps. Once inside a system, the malicious actor can collect intelligence for a deeper understanding of human and machine behaviors within a network of connected devices, such as roles and administrator privileges, passwords, communications and data flows, and additional vulnerabilities that can be exploited among the devices in the system. Such information may then facilitate the lateral movement or expanded compromise by moving to new targets within the network. The lateral movement can facilitate malicious actor activities such as refined targeting or obfuscating the attack.

Command, control and actions on objectives occur when the malicious actor has established itself within the internal network, and potentially created a connection to an outside server. The malicious actor does this for the purposes of adapting actions to achieve end objectives. Such actions could include stealing or degrading data, or manipulating system functions.

Exfiltration and sanitation, which are the final steps of compromise, is where the malicious actor removes themselves from the targeted devices and network, and potentially attempts to remove any indicators that they have been compromised in the first place.

4.3 Technology readiness

With technology readiness, this report relies on a simplified version of the classic technology readiness level (TRL) scale. The TRL scale typically consists of nine levels, starting with low maturity at the first level and ending with high maturity at the ninth level. At the first level, basic principles about a concept have been identified, whereas at the ninth level, a fully commercialized system exists (NASA, 2020). In the simplified version used for the purposes of this report, the levels of the scale are divided into idea, prototype, validation, and production (CloudWatch, 2016), as explained in table 3. Moreover, this

report also considers whether malicious use of the technology has been demonstrated or not in real cases of cyberattacks.

Table 3 Simplified technology readiness levels

Production	The product has been commercialized and exists on the market (or has been produced and used by a malicious actor).	TRL 9
		TRL 8
Validation	A prototype is tested and assessed in a realistic environment.	TRL 7
		TRL 6
Prototype	A prototype or demonstrative experiment has been made and tested.	TRL 5
		TRL 4
Idea	An idea of a technology or its use exists and fundamental principles relating to the use case are identified.	TRL 3
		TRL 2
		TRL 1

5 Previous AI research at FOI

Previous research on AI at FOI has focused on the wider implications of technology development. Strömberg (1987) published the first FOI technical prognosis for AI technology development. Since then, some of the predominant areas of AI study at FOI has included AI for intelligence use, planning, and decision support systems in military and non-military contexts (e.g. Schubert, 2017; Brynielsson et al, 2018). FOI has also conducted related research on the application of AI autonomous platforms (e.g. Svenmarck et al, 2020), and autonomous weapons systems (e.g. Hagström, 2016).

In a more recent technical prognosis of AI, Gisslén (2014) made an overarching mapping of AI-capabilities (e.g. autonomous robotics, text analysis, image analysis, etc.) as well as covering some of the international security policy dimensions of the technology. Notably for this report, Gisslén notes that future botnets could be augmented with (AI-based) swarm technology and that computer viruses could be combined with various AI technologies. Svenmarck et al (2020) similarly note that AI could be used to augment digital intrusions in unmanned aircraft system networks. Zouave (2019) noted the overlap in international efforts to develop international law norms concerning autonomous weapons systems and cyber operations.

In related, but not necessarily AI-oriented research, Holm and Sommestad (2016) explore the automation of attack plans in cyber security exercises using the tool SVED (Scanning, Vulnerabilities, Exploits and Detection). These efforts strive for reliability and replicability in cybersecurity experiments. In addition, SVED can run distributed automatic execution of sequences and record the different activities within an action. The article tested SVED through an example experiment and found it successful in terms of the scale of the experiment. The research facilitates the automation of cyber security training exercises in part.

Similar to the research on SVED, Löfvenberg, Sommestad, and Bildsten's (2019) surveyed the current state of a new category of software programs designed to detect, analyse, and exploit software bugs with the goal of conducting cyberattacks on computer systems automatically. The report illustrates that the field is still in its infancy, with almost all of the surveyed programs capable of only simplistic attacks on weakened targets using previously discovered vulnerabilities. As a result, the literature review finds that the attacks these automated systems produce are not effective against the software and defences prevalent in today's computer ecosystems. However, Löfvenberg et al (2019) expect this technology to mature rapidly in the years to come.

6 Results: AI-supported cyberattack anatomy

On the basis of 96 sources found in the literature review, this section identifies 19 use cases for artificial intelligence in the cyberattack anatomy. The use cases and their placement in the stages of the cyberattack anatomy is illustrated in table 4. The results section of this report has been structured according to these stages and use cases. The literature was further processed to identify predominant AI-techniques used and blogs, industry, and media sources were processed to find examples of real cases of malicious use of the technology.

Table 4 Found use cases for AI in the stages of the cyberattack anatomy

Reconnaissance	**Access and penetration**	**Internal reconnaissance and lateral movements**	**Command, control, and actions on objectives**	**Exfiltration and sanitation**
Strategic intelligence collection Target profiling Vulnerability detection Outcome prediction	Attack planning Phishing and spear phishing Attack code generation Classifier manipulation Password attacks Captcha attacks	Network and system mapping Network behavior analysis Smart lateral movements	Domain generation Self-learning malware Swarm-based command and control NLP manipulation	Discovery obfuscation “Low-and-slow exfiltration”

The amount of sources processed for each stage in a cyberattack in his report is shown in table 5, below. The table only shows the actual sources identified as relevant for the results of the report. It only accounts for research literature, thus not counting news, blogs and other media.

Table 5 Research publications processed for each stage of the cyberattack anatomy

Reconnaissance	**Access and penetration**	**Internal reconnaissance and lateral movements**	**Command, control, and actions on objectives**	**Exfiltration and sanitation**
39 sources	24 sources	14 sources	14 sources	5 sources

6.1 Reconnaissance

AI generally provides a capability to efficiently process textual data at a rate that otherwise would be too strenuous, if not impossible for humans. This makes AI technologies a powerful tool for reconnaissance; both in its aggregated and targeted forms. In their study on semi-automatic data-driven web analysis, Rosell et al (2018) note the wide potential of rule-based methods of web analysis for intelligence purposes. They explore both text analysis and image analysis for intelligence purposes. Guarino (2013) proposed two reconnaissance related uses of AI in his early study on autonomous intelligent agents in cyber offence, namely vulnerability discovery and profiling. The sources reviewed for this report particularly highlight four use cases for AI-supported reconnaissance in cyber operations; strategic intelligence collection and analysis, target profiling, vulnerability detection, and outcome prediction.

6.1.1 Strategic intelligence collection and analysis

Strategic intelligence are reconnaissance outputs that support the planning and policy formulation associated with cyberattacks. This may include evaluating potential types of targets, assessing tools, techniques and procedures, mapping known countermeasures, and understanding perceived ramifications and potential responses from targets (e.g. Perera et al, 2018). The predominant AI-technology employed throughout the literature is Natural Language Processing (NLP), which is common for analyzing written human language in digital form.

The data collection and processing can span a range of web-based sources. For example, studies by Mulwad et al (2011), Jones et al (2015), Perera et al (2018) and Juric et al (2019) used natural language processing (NLP) on open sources materials such as Twitter, forums, thematic databases, and other online forums. Similarly, Phandi et al (2018) applied natural language processing (NLP) to an extensive database of malware reports. Comparable tools for cyber security professionals, such as TheHarverster and Spiderfoot, perform open source (e.g. Shodan and HaveIBeenPwned), and social media scanning (e.g. Twitter, Facebook and Instagram) as well as scanning of restricted online sources (e.g. darkweb) (Kali Tools, 2019; Spiderfoot, 2019). Some concrete examples of information that might be collected this way include (Kali Tools, 2019; Juric et al 2019; Spiderfoot, 2019; Dheap, 2017):

- public analyses of cyber security trends,
- incidents and suspected events,
- vulnerabilities and exploits,
- common security practices and security measures (e.g. within industries),
- organization names, business activities, locations and opening hours of potential target organizations,
- employee names, emails addresses, social media user names, and social networks,
- host addresses and ports, domain hierarchies, etc.

While natural language processing (NLP) is fundamental to the literature reviewed here, it was further combined with other technologies and methods. Juric et al (2019) employed semantic web to make internet data understandable to machines. These technologies include conceptual modelling of the data, such as vocabularies, through resource description framework (RDF), semantic class hierarchies through web ontology language (OWL) as well as semantic web rule language (SWRL), which expresses rules and logic. Juric et al (2019) used keywords such as “vulnerability” and “exploit” as well as Big Data technology, IBM Watson Analytics, and ReactiveX to make sense of tweet semantics and keyword relations. Perera et al (2018) used natural language processing (NLP) for (linguistic) event detection in heterogeneous online textual data to categorize these according to the cyber kill chain. They use kill-chain related keywords and flexible parsing to distinguish syntax and relevance. Probabilistic soft logic (PSL) is used together with cascading hidden Markov models (cascading HMM) to detect relations in semantic data and to make predictions about future cyberattacks based on the data. Mulwad et al (2019) relied on a support vector machine (SVM) to classify vulnerability descriptions, Wikitology for topic identification and extraction, and semantic OWL to derive machine understandable assertions. Phandi et al (2018) analyzed annotated cyber security reports with machine learning such as support vector machine (SVM), conditional random fields (CRF), and naïve Bayes (NB) for classification and pattern recognition. Dheap (2017) suggests NLP and web crawling to surveil social media and cyber security bulletins. He makes the prediction that various types of AI techniques can be combined to facilitate cognitive cyber threat hunting in the near future. Dheap (2017) further notes that natural language processing (NLP), and neural networks (DNN) are current applications for AI-supported threat and risk research. These analyses should be considered dual use (both defensive and offensive) in as much as they automate general information collection on specific types of attacks, and specific factors that impact risks.

The findings of this study suggest that AI-augmented strategic intelligence is at a high technology readiness level. Machine learning is currently being implemented in a growing number of threat hunting and threat analysis products. Vähäkainu and Lehto (2019) provide an excellent overview of such technologies. In one of their examples, antagonists could potentially use AI-supported analysis of known malicious software signatures to avoid the known malicious behavior, such as is implemented in CylanceProtect (Vähäkainu & Lehto, 2019). Furthermore, a number of threat hunting and threat detection platforms on the market are already implementing machine learning to correlate network behavior to incident data, such as the Vectra Cognito Platform (2019) and Hunters.AI (2020). More well resourced antagonists might attempt to study, learn from, and copy defensive machine learning products in the future, especially as the number of solutions increase in accessibility.

6.1.2 Target profiling

AI has already had demonstrated effects on the ability to profile information and communications technology use. AI is currently used by major social media and e-commerce platforms for profiling and targeted advertising and has, in some cases, increased revenue manifold as a result (Brynjolfsson & McAffee, 2017). Bilal (et al 2019) provide a general taxonomy of profiling and the AI methods that support these methods. Bilal et al note that there are two overarching types of profiling, the profiling of individuals and the profiling of groups, and that machine learning, convolutional neural networks (CNN), and fuzzy logic ontology are the predominant AI-methods employed.

This report considers both profiling against persons (or groups of persons), as well as machines and networks. Brundage et al (2018) presents the scenario of cyberattack target profiling based on their social media content, noting that:

> Public social media profiles are already reasonably predictive of personality details, and may be usable to predict psychological conditions like depression. Sophisticated AI systems might allow groups to target precisely the right message at precisely the right time in order to maximize persuasive potential.

In this regard, Kirat et al (2018) propose that AI agents could profile targets to increase the likelihood of success. Dheap (2017) suggests that deep neural networks and neural models can currently be used for target classification and profiling.

The results from this report's literature review suggests that to date, technical research from the antagonist perspective on profiling is still relatively rare. Bahnsen et al (2018), also confirm this in their research on malicious machine learning (ML). Zhou (2018) and Bahnsen et al (2018) both note the malicious potential in technology such as Honey-Phish, using Markov chains for automated text message creation in response to phishing attempts ("scam baiting"). They similarly highlight the use of Markov models and long term-short term memory (LTSM) neural networks in SNAP_R (see section 2.1 of this report). In their own research, Bahnsen et al (2018) rely on recurrent neural networks (RNN) and long term-short term memory (LSTM), not to profile individual users or accounts through patterns in their communication, but to learn patterns in successful synthetic URL-generation. Seymour and Tully (2017) apply machine learning, specifically affinity clustering, spectral clustering, balanced iterative reducing, and clustering using hierarchies (BIRCH). They do this to evaluate potential phishing targets on social media based on whether they are likely to respond to a phishing attempt, or whether they are a high value target (e.g. CEO of an organization). Identified targets also had their posting behavior profiled for topics. Topical content profiling was realized through dictionary frequency and stop word, although Seymour and Tully (2017) also considered using a Hidden Markov Model (HMM) for their experiment.

Conversely, one of the key findings of our literature review is that the identified research on profiling in the cyber security domain is predominantly defensive, albeit the technology may be dual use. These identified areas of research include adversary profiling (Brynielsson et al, 2016), attacker profiling (Kapetanakis et al, 2014; Filippoupolitis, Loukas & Kapetanakis, 2014; Al Fahdi, Clarke & Furnell 2013), and attack profiling (Yarng, Ray &

Maher, 2003). While this research is not the focus of this paper, a couple examples are nevertheless worth mentioning. For example, in their inventory of emerging hacker assets, Samtani et al (2017) test social network analysis (SNA) to identify content authors, produce relational metrics between online accounts and content authors and map data points of interests, such as "technology diffusion between individuals". Brynielsson et al (2016) suggest training profiling systems by creating profiled personas during cyber defense exercises. However, it should be reiterated that Brynielsson's work specifically focuses on profiling the attacking side (rather than defenders), such as by classifying motivation (e.g. spy, insider, or ideologue).

The successful application of AI for profiling in cases such as CyberLover and SNAP_R demonstrate a relatively high readiness level of these technologies. With respect to these cases, Zhou (2018) remarks that, "[s]ince we know that similar models are already deployed […] being cautious on traditional phishing platforms like email is not enough." Moreover, additional AI-applications with dual use potential already exist. It is foreseeable that AI-supported reconnaissance tools will fall on a spectrum from simple to sophisticated. Simple tools process homogenous data on targets from relatively few sources whereas sophisticated tools are capable of dealing with heterogeneous types of data, and combining (and possibly evaluating) the data for more detailed profiles, potentially on multiple targets at a time. An example of a relatively simple tool that demonstrates this capability is ExifTool, which facilitates reconnaissance on, inter alia, image, audio and video metadata (Harvey, 2019). Examples of tools that demonstrate capabilities that are more sophisticated are TheHarverster and Spiderfoot, producing analytics and visualizations based on data and sources such, as (Kali Tools, 2019; Spiderfoot, 2019):

- Email addresses,
- Employee names and user names,
- Hosts,
- Port scanning,
- Subdomains,
- Social media scanning (e.g. Twitter, Facebook and Instagram),
- Open source scanning (e.g. Shodan and HaveIBeenPwned),
- Darkweb scanning.

Through research prototypes such as SNAP_R (Seymour & Tully, 2016) and actual malicious profiling seen in CyberLover (Rossi, 2007) it is possible to conclude that both the technology readiness level of this technology is high as well as the likelihood of malicious end use. Considering CyberLover, it should be noted that natural language processing (NLP) of social media is the most evident profiling use case identified among malicious actors to date.

6.1.3 Vulnerability detection

In last year's literature review on automatic attack code generation (Löfvenberg, Sommestad & Bildsten, 2019), the authors came to the overarching conclusion that the technology was still "relatively immature". The solutions found in the study primarily focused on the detection of previously known and common vulnerabilities. As Löfvenberg, Sommestad and Bildsten (2019) focused on literature that demonstrated the exploitability of detected vulnerabilities, this report primarily review at complementary literature. The academic research found in this report spans over a period of 15 years, from Corral et al (2005) to Juric et al (2019). The identified literature approaches the subject of vulnerability through textual analysis of online sources and vulnerabilities associated with activities in web browsers. The applied AI techniques generally involve regression and classification methods.

While focusing on results from more recent years in this report, it is noteworthy that Mulwad et al (2011) presents an interesting approach to gathering intelligence on vulnerabilities in web text using a support vector machine (SVM) classifier on a national vulnerability database. More recent studies (Almukaynizi et al, 2017), uses machine learning to generate exploit predictions based on online vulnerability mentions. Almukaynizi et al (2017) concluded that their experiments resulted in high true positive rates (90 percent) and low false positive rates (less than 15%) for exploit prediction. They experimented with several machine learning techniques, including SVM, random forest (RF), naïve Bayes (NB), logistic regression (LOG-REG) on vulnerability metric features determined through an open vulnerability scoring framework known as Common Vulnerability Scoring System (CVSS). Zhang, Ou and Carragea (2015) attempted to apply a machine learning (ML) model on a national vulnerability database (NVD) to predict zero-day (unknown) exploits, finding that the NVD was generally not conducive to such predictions. However, they found that some of their models were more promising than others; for example one was capable of predicting zero-days in Firefox and Internet Explorer. Zhang et al (2015) used seven different regression functions in their experiment and six different classification functions.

One of the areas of research found in this report focuses on the detection of technical vulnerabilities associated with activities in a web browser. Almousa and Anwar (2019) propose to predict the risk that websites will exploit browser vulnerabilities through feature extraction with feed forward neural networks, convolutional neural network (CNN), and a classifier model. The ambition is that the model will be able to classify websites as benign, suspicious or as exploit websites. Luckow, Kersten and Pasareanu (2020) designed a classifier based on logistic regression to detect vulnerabilities derived from algorithmic complexity. While they note that the classifier could also have relied on artificial neural networks (ANN) or SVM, their model was able to detect 87.5 percent of the identified vulnerabilities in the test. Dhaya and Poongodi (2014) created a system to detect vulnerabilities in mobile phone applications, which tend to be unverified by the authorizing company (i.e. Android), with an N-gram analysis classifier, also based on the CVSS. They conclude that the system is effective but does not identify previously unknown vulnerabilities.

Löfvenberg, Sommestad and Bildsten (2019) considered 17 identified solutions in their literature review on attack code generation. This report does not find any cause to revise their finding that while prototypes have been demonstrated in the research literature, those prototypes are generally not sophisticated. It should be noted that the literature in this report is primarily concerned with automated detection of *known* vulnerabilities. Moreover, as there are market solutions such as Sovereign Intelligence (Schroer, 2020; Sovereign Intelligence, 2020), which utilizes AI-supported analyses of open source, dark web, deep web and peer-to-peer data to predict vulnerabilities, this technology must be considered to have a high technology readiness (see also ImmuniWeb, 2020).

6.1.4 Outcome prediction

Dheap (2017) suggests the utility of machine learning for outcome prediction in the near future. He suggests that AI tools can analyze current and historical events to predict the results of future planned actions, also leading to applications where AI could recommend future actions. The development of assessment and simulation methods related to cyber operations could be vital steps on the way towards more advanced outcome prediction models. Specifically for the malicious actors, Dheap suggests that developments in AI could bolster their confidence to seek riskier, more high value outcomes when this is made possible through AI-supported attacks. The literature review of this report only demonstrates experimentation with specific instances of evaluative and predictive AI. Examples include evaluative profiling of targets to assess the likelihood of successful phishing (Seymour and Tully, 2017) (see section 6.1.2), vulnerability prediction (Zhang et al, 2015; Almukaynizi et al, 2017)(see section 6.1.3), exploit risk prediction (e.g. Almousa & Anwar, 2019), as well as assessments relating to attack planning(see section 6.2.1).

6.2 Access and penetration

This section of the report summarizes the findings on AI technologies augmenting the planning and execution of attacks by which malicious actors can gain access to target systems and networks. In a 2013 study (Guarino, 2013), it was proposed that AI would be used in offensive cyber operations to generate attack plans. This findings in this report, based on 24 sources, confirm the viability of attack planning, phishing and spear phishing, attack code generation, classifier manipulation, password attacks (i.e. guessing, brute forcing and stealing, as well as captcha attacks as viable uses cases for AI. Moreover, it is further found that commercialized solutions for defeating captchas utilize machine learning and that such attacks have been observed in the wild.

6.2.1 Attack planning

Generating attack plans is an established practice within network security and is a way to potentially understand the effects vulnerabilities may have (Randhawa et al, 2018). An attack plan could be defined as a "sequence of actions which, if taken by a person or computer, could harm the target organization" (Yuen, 2013). The automated planner of an attack is a branch within AI and attack planning can be considered time consuming for human users (Yuen, 2013). An automated attack planner is a method within artificial intelligence that can support the decision and deliberation process of attack planning. Further, attack planning is related to decision theory (Ghallab et al, 2004). An automated planner of an attack also increases the accuracy and comprehensiveness of the assessment (Yuen, Turnbull, & Hernandez, 2015). Researchers have, therefore, tried to generate new methods of using AI in the initial planning stage of attacks in order to increase the effectiveness. For this reason, automated Cyber Red Teaming is used as an exercise to ascertain viable attack plans (Yuen, 2013).

Randhawa et.al. (2018) presents a system application (Trogdor) that uses multiple AI planners to perform vulnerability assessments automatically. This is done by generating several attack graphs that are able to target vulnerabilities and critical structures. The application makes use of "a library of Tactics, Techniques and Procedures" that can model behaviour of the target system and reveal its vulnerability structure (Randhawa et al, 2018). Further, the use of Trogdor to generate attack graphs enables the analyst to prioritize actions to the most critical resources in the target systems (Randhawa, et al, 2018).

Attack graphs are used to establish the most beneficial attack path for a malicious actor. However, pathways vary between the least complicated route and the most advantageous attack path, which may be ascertained by quantification metrics (Falco et.al. 2018). The effectiveness of an attack tree was measured as the time it took to build the attack tree. Automated creation of attack trees was considerably faster than manually building attack trees. Further, the automated attack tree can be standardized and will therefore not require high-degrees of cybersecurity knowledge (Falco et al, 2018).

6.2.2 Phishing and spear phishing

AI-augmented phishing and spear phishing are a form of automated decision-making facilitated by prior target profiling. Several sources (Seymour & Tully, 2017; Zhou, 2018) refer to the HoneyPhish project, which used a hidden Markov Model (HMM) to automatically generate response texts to phishing emails from scammers. However, Seymour and Tully (2017) notes that the system did not yield believable English language sentences, and thus had a low response rate, meaning that few scammers fell for the AI-generated bait. Instead, they designed SNAP_R, using a recurrent neural network (RNN), Markov models (MM), and long short-term memory (LSTM) to generate spear phishing tweets with fraudulent links. They report that the experiment generated a click rate between 30 percent and 66 percent. One of the obstacles to successful phishing campaigns is bypassing automated phishing and spam detection technologies. The research of Bahnsen et al, (2018) focuses on the malicious use of machine learning (ML) to bypass AI-based

phishing detection systems. Specifically, their DeepPhish algorithm used a recurrent neural network (RNN) and long short-term memory (LSTM) classification that generated fraudulent URLs which could potentially be used in a phishing scenario and that can avoid automated detection. Bahnsen et al (2019) find that DeepPhish increased the success rate of detection bypass as compared to manual attempts.

According to Rossi (2007), the CyberLover malware used natural language processing (NLP) to generate dialogue in online chat rooms. CyberLover profiled individuals seeking relationships (e.g. as "romantic lovers" or sexual predators) to steal personal data and automatically generate customized messages containing malicious links and attachments.

6.2.3 Attack code generation

As previously mentioned, this report was preceded by an FOI literature review on automatic attack code generation (Löfvenberg, Sommestad & Bildsten, 2019), looking at the research literature covering automatic vulnerability detection and exploitation. Their literature review identified 22 publications where the research proved that detected vulnerabilities were exploitable and where generated attack codes could be used to leverage administrative privileges on the target computer. Within the 22 publications, they account for 17 different solutions. Löfvenberg, Sommestad and Bildsten's (2019) overall assessment was that identified tools were relatively unsophisticated and primarily focused on well-known and old vulnerabilities. While some of their identified solutions were being commercialized, they determined that extensive work would be needed to create a qualified solution capable of generating attack codes against real software. Moreover, they deem that two strategic interests in this technology include defensive prioritization of vulnerability patching, as automatically exploitable vulnerabilities ought to be prioritized in a future of automated attack code generation, as well as offensive mass-generation of attack codes, allowing attackers to stay ahead of defenders. This report does not replicate the methods of Löfvenberg, Sommestad and Bildsten, (2019), but reviews some of the most recent literature published in 2019 and later.

Scanning the 2019 literature regarding AI-augmented generation of conventional exploits, only one new source was found. Wang et al (2019) do not use artificial intelligence in their experimental exploit generation focusing on the internet of things (IoT). Instead, they design a genetic algorithm, based on an artificial bee-colony algorithm and simulated annealing algorithm, for scheduling to improve the overall efficiency of automated vulnerability detection and exploit generation.

Looking beyond pre-AI vulnerabilities and exploits, there is tendency that some researchers (San Agustin, 2019; Moisejevs, 2019; Chakraborty et al, 2018) characterize security flaws in machine learning, deep learning, and classifiers (see next section) as a form of vulnerability that attacks can exploit (San Agustin, 2019; Moisejevs, 2019). Chakraborty et al (2018), provides a taxonomy of several such attacks, notably generative adversarial attacks, adversarial example generation, generative adversarial network (GAN) based attacks in collaborative deep learning, adversarial classification, evasion and poisoning of support vector machine (SVM), poisoning of collaborative systems, and adversarial attacks on anomaly detection. Given the increasing use of potentially vulnerable AI (Chakraborty et al, 2018), this field of study emerges as an area for further research in automated exploit generation. In particular, there is an apprehension that such attacks may cause physical harm when directed at cyber-physical systems such as smart cars (San Agustin, 2019). In this respect, San Agustin (2019) recommends adversarial training and the implementation of standardized code hardening guidelines at a global scale as countermeasures to exploitation. A similar conclusion to Löfvenberg, Sommestad and Bildsten (2019) can thus be reached; that research on AI-supported vulnerability detection and exploit generation may become a matter of priority for both defenders and attackers.

6.2.4 Classifier manipulation

Another finding of the literature review is research on attacks against AI implementation in targeted systems, namely classifiers. The use of classifiers to detect malware in supervised learning can benefit the malicious actor. Since the input data require different models of malware (Cakir & Dogdu, 2018) to label different classification entities an actor can manipulate the data purposely with adversary techniques to undermine the classifiers. This has been done in several instances by inserting fake data inputs for example or manipulating the content of spam emails in order to pass the spam filters classifications (Biggio, et.al. 2014). The naïve Bayes (NB) classifiers used for spam email filtering is done through supervised machine learning with classified data (Dheap, 2019). A malicious actor can use the classifier to defeat the spam filter by manipulating the data input and training data. In addition, it is possible for actors to recover training data via reverse examining the techniques used and then use the knowledge of the defensive techniques to defeat spam filters or anti-malware software (Truong et al, 2020).

6.2.5 Password attacks

Researchers have typically experimented with three types of AI-augmented password attacks; password guessing, password brute forcing, and password stealing. Hitaj et al (2018) present research on PassGAN, a generative adversarial network (GAN) that emulates passwords from real password leaks, demonstrate how the technology can generate password guesses that match passwords in dictionaries at higher rates than other existing password guessing solutions. Trieu and Yang (2018) note that AI and cyber security research converge to make both defensive and offensive applications smarter. They tested an open source machine learning algorithm (Torch RNN) which generates password based on patterns existing in password dictionaries. Torch RNN was thus used to brute force passwords with a success rate of 57 percent after 1000 experiments. Lee and Yim (2020) are concerned with attack techniques that steal passwords during online authentication. While noting that less advanced attack tools for key logging currently exist and are available as online assets for antagonists, they use existing techniques to train a machine learning (ML) model for the classification of data from keyboard strokes. Their model was successful in stealing keyboard data (to a 96.2 percent accuracy level). The model generated by Lee and Yim relied on "k-Nearest Neighbors (KNN), logistic regression, linear Support Vector Classifier (SVC), decision tree, random forest (RF), gradient boosting regression tree, support vector machine (SVM),and multilayer perceptron (MLP)." Further, Vijaya, Jamuna and Karpagavalli (2009) used classifiers such as decision tree, naïve Bayes (NB), multilayer perceptron (MLP) and support vector machine (SVM) to predict password strength. The work of Suganya, Karpgavalli, and Christina (2010) is likewise noteworthy using a support vector machine to the same end. While this may primarily be a defensive technology, it may also be useful to antagonists who can gain information about password quality policies in target organizations, or who can combine the technology with other methods to profile accounts, such as through websites like Have I Been Pwned. Although antagonists have been known to use automated password attacks (WSJ, 2017), this study has found no evidence of AI-augmented cases in the wild.

6.2.6 Captcha attacks

AI-based captcha attacks are a well-established area of research. Bursztein et al (2011), noted already nine years ago how developments in machine learning is causing captcha as security measure to come under scrutiny. Attacks against captchas make it possible to emulate a real human user, for example at a log-in interface in a web browser (WSJ, 2017). Approaches to AI-augmented captcha attacks vary according to the type of captcha target. Bursztein et al (2011) for example, created Deeptcha, using classifiers such as regularized least-squares classification (RLSC), and one-versus-all (OvA) classification to solve audio captchas. Their system solves 49 percent of Microsoft's captchas and 45 percent of Yahoo's, respectively. Bursztein et al (2014) use ML, notably ensemble learning, to simultaneously attack segmentation and recognition in text-based captchas. Their test solved Yahoo

captchas at a rate over five percent, ReCaptchas at a rate over 33 percent, Baidu at a rate above 38 percent, and CNN at a rate above 51 percent. Wang et al (2017) designed a convolutional neural network (CNN) and adaptive algorithm to defeat text-based captchas with multiple digits, resulting in a lowest single character recognition rate above 75 percent.

The overall research on captcha-breaking is far too extensive to describe in this report. A Google Scholar search for "machine learning defeat captcha", showing results from 2019 and forward results in 375 hits alone. Moreover, the established Death by Captcha, a captcha-solving solution, has relied on machine learning for some time (WSJ, 2017; Breck, 2020). Additionally, ticket scalpers have allegedly been able to successfully use captcha attacks against the captchas on the ticket sales website, Ticket Master (Bursztein et al 2011; McMillan, 2010; Zetter, 2010). As such, captcha attacking tools seem to be a mature technology that is available for antagonistic use.

6.3 Internal reconnaissance and lateral movement

Internal reconnaissance is the stage of a cyberattack in which the malicious actor is trying to gather internal information about the target systems and network. This is done so that an attacker can adjust their position on the network, adjust internal targeting and carry out the planned activities more effectively. In comparison with reconnaissance, internal reconnaissance occur within the target network (Metivier, 2018; Advanced Network System, 2018). Guarino (2013) alleged that AI could be used at this stage to collect process information to learn more about the network. The sources processed in this report confirm that AI can facilitate the process of internal reconnaissance to map system vulnerabilities, adapt to target behaviors and make decision based on the structure of the environment. Specifically, network and system mapping, network behavior analysis, and smart lateral movements seem like plausible applications for AI based on the findings.

6.3.1 Network and system mapping

Within the stage of internal reconnaissance, mapping and classification tools are often used to identify vulnerabilities and important information. Randhawa et.al (2018) describe how their AI-solution, Trogdor, identifies logical connections and models firewall rules on target networks. A malicious actor can use these descriptions to their advantage and map the vulnerabilities a system might possess (Randhawa et al, 2018). Another technique is topic modelling with the use of AI libraries. Topic modelling can help a malicious actor to classify and sort system applications. The malicious actor can therefore target what is seen as the vulnerability (Greeff & Ross, 2019).

6.3.2 Network behavior analysis

A system can use AI to detect abnormal behavior on the network. However, a malicious actor can adapt to the baseline of behavior that is prescribed in the systems behavioral patterns (Szmit & Szmit, 2012). With the behavioral knowledge, the malicious actor can move undetected through the system. A fundamental axioms within this area of research is that if an AI system can learn to detect sequences that relate to malware or malicious behavior another AI can learn to surpass these defensive systems (Truong, et al, 2020).

As the system reacts to changes in the conditions of the network, and detects abnormalities, a common tool for attackers to use is the shadow process. The shadow process identifies the baseline for normal behavior and then adapts the malware's movement accordingly. This is done by splitting the malware code into sequences and rewriting the malware graphs to export them into a different system processes. This process will therefore remain undetected by malware detectors since only small segments of abnormal behavior are visible (Ma et al, 2012). When an intrusion detection system has a baseline of behavior generated by deep

learning (DL), AI-supported malware can also start to impersonate the behavior of human users via contextualization (Darktrace, 2018).

Clustering algorithms are frequently used to recognize the changes in behavior. The algorithm is used to classify data into a baseline for behavior analysis in networks in order to detect anomalies with deep learning methods (Dheap, 2017, Lima et al, 2010). Deep learning methods have been applied to intrusion detection in for example deep belief networks (DBN) with various techniques of deep learning (DL). The authors try to combine several methods of deep learning (DL) for example auto encoders, deep belief neural networks (DBN), deep neural networks (DNN) and extreme learning machine (ELM) restricted Boltzmann machine (RBM). An ensemble of deep neural networks technique can, therefore, improve the overall detection rate as well as the accuracy of detection in comparison with using only one deep learning technique (Ludwig, 2017). However, adversarial attacks have optimized the malicious actor's ability to intrude on deep reinforcement learning agents. As a reinforcement learning agent interacts with its environment, the adversarial attack can correlate and produce an example of the same sequences in the environment. Thus, AI is used as both an effective adaptation on common systems, as well as an attacker's tool for malicious use (Lin et al, 2017).

The literature review indicates that AI-supported network behavior analysis is at a high readiness level and that this technology has been adopted by malicious actors for antagonistic end use. Machine learning is already being deployed in products to detect malicious behavior on networks, for example the Darktrace (2018) Cyber AI Analyst, Flowmon (2020) ADS, and the Vectra (2019) Cognito Platform. Notably, Darktrace (2018) predicts that this technology could also be leveraged by malicious actors. Allegedly, AI has already been used to this (or similar) effect by malicious actors in the case of the Morning Download (WSJ, 2017; Norton, 2017).

6.3.3 Smart lateral movements

AI-driven malware will also be able to make decisions based on the structure of the infected system in order to move undetected. This reasoning follows from the research of Truong et al. (2020), Darktrace (2018), and Lin et al. (2017). Intelligent malware will be able to "hide" within commonly used systems (for example PsExec, RDP, SSH) and be considered normal use, thus concealing itself from intrusion detection systems. Further, as seen in the research relating to network behavior analysis, the malware could also learn the infected environment by remaining quiet and thus observer the normal operations. The ability for a malware to move laterally at a higher speed is expected to facilitate infecting more devices over shorter periods of time, more autonomously (Darktrace, 2018).

6.4 Command, control and actions on objectives

In the standard cyber kill chain, command and control (C2) occurs after the install phase has been conducted (Rice, 2014). With C2 the malicious actor is attempting to establish a communications link between it and the target with the goal of exerting influence over the compromised computer system and other systems on its internal network (Wirkuttis & Klein, 2017). Depending on the motives of the malicious actor involved, a successful C2 channel can be used for a number of different purposes. These range from facilitating the spread of the malware to other networked computers, instructing the target to participate in botnet attacks, downloading and installing remote access trojans (RATs), and exfiltrating data (Rice, 2014). Pursuant to these steps, the malicious actor can begin taking actions to achieve their objectives in the system. Guarino (2013) postulated that artificially intelligent malware would be able to memorize key reconnaissance findings on the targets to leverage a more persistent presence in targeted systems than conventional malware. In this early work on the topic, he further posited that AI would facilitate malware that reacts to countermeasures and can exert decentralized controls over agents such as botnets. The sources on AI research identified for this phase of the cyberattack anatomy is primarily

concerned with autonomous C2. This research encompasses deep learning domain generation, self-learning malware and swarmed-based C2 techniques. However, the literature review also identified attacks on natural language processing (NLP) as a means to, inter alia, degrade data integrity. Moreover, this report has identified that the technology used in AI-supported domain name generation already exists as a service for security purposes.

6.4.1 Domain generation

Binary-based domain-generation algorithms (DGAs) are used by malicious actors within the C2 phase of the cyber kill chain to both establish C2 and facilitate data exfiltration (Sood & Zeadally, 2017). Their purpose is to generate thousands of pseudorandomised algorithmically generated domains (AGDs) daily, which can be used by the malware to send and receive information (Anderson, Woodbridge & Filar, 2016). Most of these queries will not succeed due to the majority of these randomly generated domain names being unregistered. However one query, which has been preselected by the malicious actor, will successfully resolve and establish a connection with the malicious actor's C2 server.

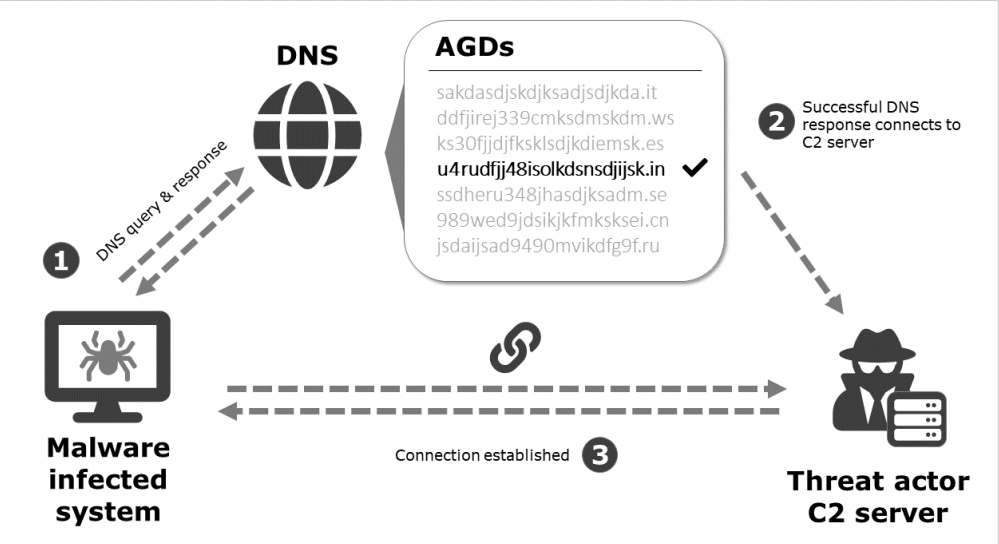

Figure 2 Binary-based domain-generation algorithms (DGA)

DGA strategies are hard to defeat without the use of modern DGA classifiers, as defenders must either "sinkhole, pre-register or blacklist all of the domains to prevent the [command and control] connection" (Anderson, Woodbridge & Filar, 2016). DGA classifiers use machine learning (ML) to examine successful DNS queries made by infected hosts and score them on various values derived from training datasets (Anderson, Woodbridge & Filar, 2016). Queries under a certain threshold are identified as DGAs and blocked.

While DGA classifiers using ML models work on simple DGAs, their use of predefined training data can be foiled by attackers using generative adversarial networks (GAN). Anderson, Woodbridge and Filar (2016) demonstrate how this is possible by constructing what they call DeepDGA, a deep learning based DGA that bypasses detection by training against itself. By using GAN, a generative model can be trained through the use of two sub-models; the generator and the discriminator. The goal of the generator sub-model is to generate domain names that the discriminator can no longer distinguish as DGA, while the goal of the discriminator sub-model is similar to that of a normal DGA classifier which must correctly distinguish between real and generated domains. By pitting these models against each other over a series of zero-sum rounds and then updating the discriminator's detection parameters after each of them, the generator sub-model gets progressively better at creating domain names that are difficult to detect. As a result of the adversarial training, DeepDGA

was able to generate domain names capable of confusing an advanced DGA classifier with manually crafted features. Akamai (Nadler, 2019), an IT-corporation, has implemented DeepDGA in their domain generation algorithm mitigation service. Akamai likewise explores whether it is possible to identify current malicious use of DeepDGA, reaching the conclusion that such uses of the technology are yet to be definitively identified.

6.4.2 Self-learning malware

Because blocking unknown incoming connections to internal hosts on a computer network is standard practice in firewall design, C2 communications often originate from the infected hosts. By using the infected host to initiate the communication, Rice (2014) argues that it would look like the host is communicating with a legitimate outside peer. Rice (2014) further reasons that if the C2 communications channel is detected it is possible to block access to it via firewall rules, thus muting the capabilities of the malware. AI-supported malware could potentially mitigate this countermeasure by giving the malware its own autonomy to act, therefore eliminating the need to 'phone home'. For example, Chung, Kalbarczyk and Iyer (2019) demonstrate, in a simulation, how a self-learning malware can indirectly attack computers in a supercomputer facility by interfering with the cyber physical systems (CPS) of the building automation system. By targeting the cooling control system managing the facility, which was assumed to be less secure than attacking the actual computer infrastructure residing there, the malware gathered the data needed for it to run through scenarios capable of disrupting cooling capability. It then implement three different attack strategies on the target autonomously to successfully disrupt it. Chung, Kalbarczyk and Iyer's (2019) simulation used k-means clustering to classify target system logical control data, and Gaussian distribution to identify attack effects on the target system. The premise behind the simulation was that once malware got into the system it had to know how it should act without any further help from the attacker. As the above example by Chung, Kalbarczyk and Iyer demonstrates, malware capable of implementing its own attack plans through a self-learning algorithm may be able to lessen the amount of knowledge required by the attacker to successfully manipulate the target system.

6.4.3 Swarm-based command and control of botnets

The master-slave relationship that characterises most C2 systems within botnets has a problem with survivability and scale. In research introduced by Castiglione et al (2014) they argue that the former robustness problem is due to the increasingly sophisticated detection and mitigation techniques, while the latter scalability problem relates to the rise in both the elements to be controlled by C2 and the increasing heterogeneity of communication infrastructure. One potential way these researchers propose to get around such difficulties is by implementing decentralised swarm-based intelligence within the botnet itself. Their methodology is inspired by how ants optimise their foraging behaviour through the use of pheromones for indirect and asynchronous communications between agents. They reason that if a botnet could operate under the same mechanisms, the "only communication channel needed would be the one used to leave traces (and detect traces left by others) in an environment" (Castiglione et al, 2014). This creates a self-organizing channel in the environment, where agents obtain information from the traces to understand what their next action should be.

Castiglione et al. argue that in practice this is achieved by the node (i.e. bot) sending out a periodic multicast 'heartbeat' network packet after it joins the botnet. Older nodes already in the mesh network then learn of the presence of this new member from its heartbeat and update their topology dynamically through a degree-constrained minimum spanning tree (DCMST) algorithm. These botnet nodes then established preferred routes to each other within the network based on the principle of least cost routing, which functions similarly to how ants define their preferred path of travel by routing through the area with the highest pheromone strength. The advantage of such a topology gives the bot master the ability to cryptographically sign commands that can then be sent to one or more nodes within the

botnet at random. Received commands are decrypted by the targeted node, actioned, and then propagated onward to neighbouring nodes who repeat the entire process. Upon experimenting with their methodology, the authors' results found that a self-organizing C2 architecture could be used to enhance a botnets survivability and scale by allowing it to dynamically adapt to changing network conditions on-the-fly.

On the theoretical level, Danziger and Henriques (2017) propose a botnet framework where bots use machine learning (ML) techniques to attack a target. This framework has no C2 capabilities, removing the ability for the bots to phone home after being deployed. Instead the bot master permeates each individual bot with various machine learning (ML) methods, roles and data sets in the pre-deployment stage, which are then utilised autonomously after infecting a device. The authors note that such a framework would undermine network analysis techniques to detect botnets.

Finally, swarm-based botnet techniques involving C2 have also been mentioned by Kubovič, Košinár & Jánošík (2018). Similar to Castiglione et al. (2014), these authors suggest AI could be used to allow bots to learn and share information between each other. However, what differentiates the authors is that Kubovič et al. propose that the bots be used to run penetration tests on a target, with each individual bot being given a separate method to infiltrate the target. The results would then be reported back to the bot master who could use this information to carry out a more extensive attack.

6.4.4 NLP manipulation

Natural language processing (NLP) has lately been using deep learning and neural networks as it opens up for more accurate and more effective processing (Dheap, 2018). There are a wide array of neural network that can be used for natural language processing (NLP), including time delay neural networks (TDNN), convolutional neural networks (CNN) and Recursive Neural networks (Fahad and Yahya, 2018). While deep neural networks (DNN) have been used in NLP, Zhang et al, (2019) account for attacks using adversarial examples on these. Adversarial attacks on text data, such as gradually changes to invalid words or changes of the word sequence, can ultimately alter the semantic meaning of the text (Zhang et al, 2019). Advances in generative adversarial networks (GAN) can also impact these types of attacks, a technology which can automatically produce adversarial samples (Truong et al, 2020).

6.5 Exfiltration and sanitation

The literature review indicates that research on AI exfiltration and sanitation is not prolific within academia at present. However, a number of authors comment on the anticipated future use of AI for these means. The five relevant sources found on the topic has primarily been identified within reports on tests and investigations by the IT-security industry, and concerns the topics of pre-emptive discovery obfuscation and "low-and-slow" data exfiltration.

6.5.1 Discovery obfuscation

In 2018, IBM Research developed an AI-supported malware called DeepLocker to show how current malware technology can be merged with existing deep learning techniques to create a new breed of malware (Kirat el al, 2018; Security Intelligence, 2018). What makes DeepLocker unique is its ability to hide in plain sight by embedding itself in carrier applications like video conferencing software using an encrypted payload. This payload is undetectable by most antivirus software due to its encryption. It is only decrypted and triggered once the malware's deep neural network model has identified its specified victim via advanced trigger conditions like facial recognition (Security Intelligence, 2018). Similar to a zero-day exploit, where the vulnerability used as the attack vector was previously unknown, an infected machine with a dormant encrypted payload has the advantage of not being easily countered until after the attack is actually detected. From a sanitation

perspective, the malware's concealment capability might have implications for detection by forensic examination. Without the ability to know if a target is infected, the propensity disposed toward believing a functional network is secure will be undermined. Furthermore, Kubovič, Košinár, and Jánošík (2018) argue that malware can become aware of the environment it was in and alter its behaviour to act benign or even self-destruct when it sensed something suspicious. While sandbox-evading malware already exists today, their evasion tactics rely on pre-programmed software calls which can be outsmarted if the malware researcher knows what to look for. With AI-supported malware this may not be as easy as the malware may actively counter the researcher based on their inputs similar to how deep learning models operate to counter human opponents in popular strategy games like StarCraft II (DeepMind, 2019).

6.5.2 "Low-and-slow" exfiltration

In a research white paper by cyber security firm Darktrace (2018), it is suggested that the data exfiltration method called 'low-and-slow' would be made significantly harder to detect with AI. The 'low-and-slow' technique transfers small quantities of data over a duration of time while the target is largely unaware of the process. Darktrace (2018) argues that a contextually aware AI malware would be able to assess the target machine's pattern of bandwidth usage and make data exfiltration faster by piggybacking upon existing high-throughput activities such as video conferencing. Furthermore, as argued by Stapleton and Stevens (2019), a contextually aware malware would be able to blend in with the environment it finds itself in which has repercussions for detection and exfiltration.

6.6 Limitations to an AI-supported cyberattack anatomy

Much of the reviewed literature has not provided concise discussions on the potential limits to the AI applications. Notably, the sources assume that the readers can determine for themselves whether AI is actually needed (as opposed to less sophisticated automation) to achieve the practical objectives in the research design. Provided the extensive scope of this study and the 19 use cases identified in it, making such a determination on a case-by-case basis would be impractical. The found uses cases can therefore merely be regarded as an indication of a present trajectory in the research and not as a prediction of present or future malicious AI use. In this regard, Dheap (2019) refers to the need to determine where malicious actors will find value for the technology in the dynamic cyber threat and defence landscape. Moreover, the literature identifies general limits to the application of AI. Firstly, the quality of machine learning outputs are dependent on the quality of data inputs (Brynielsson et al, 2018; Dheap, 2017), both in a training phase and in an application phase. This would mean that defensive, dual use and offensive technologies are more likely to develop faster where there are high quality datasets readily available. Arguably, household technology (e.g. smart phones and computers) and household digital traces (e.g. on social media) become primary targets of this technology development due to their accessibility. Conversely, the sources processed during this study indicate that training AI to recognize the unknown, such as zero-day vulnerabilities, is more challenging. Wirkuttis and Klein (2017) also note that AI systems that make pattern-based predictions tend to be challenged by dynamic and evolving problems. Secondly, the black box characteristics of AI, specifically the inability of humans to understand the underlying logic, can obscure errors in automated analyses and decision-making (Dheap, 2017). This creates a trust problem where decision-makers (possibly even malicious ones) must consider whether they can adequately trust AI-solutions enough to use them (Brynielsson et al, 2018). Brynielsson et al (2018) consider the developments in explainable AI (XAI) to this effect. The ease with which cyberattacks are successfully conducted should not be overestimated. In some systems, it may be a considerable challenge for malicious actors to predict how to achieve a desired effect, even without the complexity of having to design an AI for this purpose. Finally, AI can itself be subject to attacks and manipulation (Dheap, 2017; Chakraborty et

al, 2018; Gu, Golan-Davitt & Garg, 2019) such as by inserting flawed data in the training of the AI or foiling the classifiers. Imagine, for example, a future of misinformation implanted in strategic data resources, such as vulnerability databases and the effects it might have on the training of offensive tools.

7 Discussion on findings

The results suggests that open source materials on AI applications are currently security-focused and written from the perspective of defenders, rather than offense-focused and written from the perspective of the antagonist. The report nevertheless identified 19 use-case for AI within cyberattack anatomy on the basis of research explicitly taking the antagonists' perspective or research where proposed applications have apparent dual use potential. Generally, earlier phases of the cyberattack anatomy have a higher technology readiness level than the later ones. In particular, strategic capabilities augmented by mature AI in the use cases are notably data aggregation, repetition, deception, and manipulation. Such strategic capabilities seem particularly appealing to malicious actors. However, it should be noted that to date, very few AI-augmented cyberattacks have been identified. Moreover, this antagonistic use cases found in this report demonstrate two possible scenarios; AI versus human technology user and AI versus technology. In the case of AI versus technology the literature review also demonstrates dimensions of an arms race between AI-supported attackers and AI-supported defenders.

7.1 Technology readiness

The development of certain AI technologies and the discrepancies in experimental research indicates that some use cases for AI within the anatomy of cyberattacks are currently more mature than others. The found use cases generally demonstrate a higher technology readiness level at the earlier stages of the cyberattack anatomy, especially from the reconnaissance phase to the internal reconnaissance and later movement phases, as illustrated by figure 5. Figure 5 demonstrates a mean technology readiness across use cases within phases in the cyberattack anatomy.

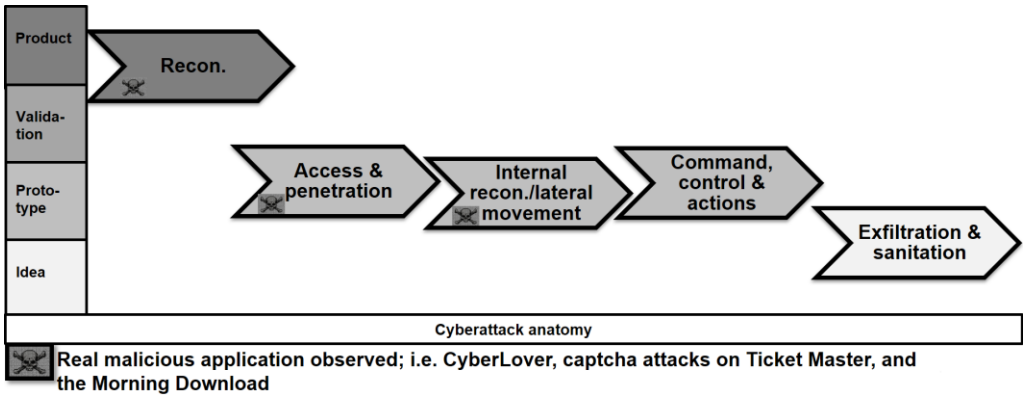

Figure 3 Overall technology readiness level for AI use in the cyberattack anatomy[2]

The overall findings on the use cases are broken down in table 6 (below), which also highlights cases of alleged real use by malicious actors in grey. At the reconnaissance phase, several of the identified use cases are already in production within the security industry. For example, malicious target profiling using natural language processing (NLP) has already been observed in the CyberLover attack. Web scanning for strategic intelligence has already been rolled out in production for cyber security, and can be adopted by malicious actors. In the case of vulnerability detection, while being at a high technology readiness level, the research findings do not demonstrate high levels of sophistication among existing products. At the access and penetration stages to the internal reconnaissance and lateral movement, research prototypes typically exist. AI-supported cyber security solutions for network

[2] Skull from Shutterstock: https://www.shutterstock.com/image-vector/skull-crossed-bones-danger-piracy-sign-550387903?irgwc=1&utm_medium=Affiliate&utm_campaign=Pixabay+GmbH&utm_source=44814&utm_term=https%3A%2F%2Fpixabay.com%2Fsv%2Fimages%2Fsearch%2Fskull%2520computer%2F

behavior analysis already on the market and that similar technology has allegedly been leveraged in real attacks. Conversely, the literature implies that AI-supported exfiltration and sanitation has fewer identified use cases for AI-supported and that these are less matured on average. In fact, the literature review did not show any explicit research into AI-supported sanitation. The approach proposed by the research rather reflects antagonistic strategies to obfuscate attack discovery, potentially also making ex-post forensics more complicated.

Table 6 Found use cases for AI in the anatomy of cyberattacks

Step in the anatomy	AI use case	Technology readiness	Malicious use
Reconnaissance	**Strategic intelligence collection and analysis**	In production for security purposes	Not demonstrated
	Target profiling	Produced by malicious actors	Demonstrated *(CyberLover)*
	Vulnerability detection	In production for security purposes	Not demonstrated
	Outcome prediction	Idea (specific types of predictions have been demonstrated)	Not demonstrated
Access and penetration	**Attack planning**	Prototype made by researchers	Not demonstrated
	Phishing and spear phishing	Produced by malicious actors	Demonstrated *(CyberLover)*
	Attack code generation	Prototype made by researchers	Not demonstrated
	Classifier manipulation	Prototype made by researchers	Not demonstrated
	Password attacks	Prototype made by researchers	Not demonstrated
	Captcha attacks	In production for security purposes Also produced by malicious actors	Demonstrated *(Ticketmaster attack)*
Internal reconnaissance and lateral movement	**Network and system mapping**	Prototype made by researchers	Not demonstrated
	Network behavior analysis	In production for security purposes Also produced by malicious actors	Demonstrated *(the Morning Download)*
	Smart lateral movements	Idea	Not demonstrated
Command and control	**Domain generation**	In production for security purposes	Not demonstrated
	Self-learning malware	Prototype made by researchers	Not demonstrated
	Swarm-based command and control of botnets	Prototype made by researchers	Not demonstrated
	NLP manipulation	Prototype made by researchers	Not demonstrated
Exfiltration and sanitation	**Discovery obfuscation**	Prototype made by researchers	Not demonstrated
	"Low-and-slow" exfiltration	Idea	Not demonstrated

7.2 From research to malicious end use

During the research presented in this report, very few confirmed cases of actual AI-supported cyberattacks were identified, even though certain aspects of the technology is mature. The automated profiling and phishing of CyberLover (Rossi, 2007) from 2007 is the earliest identified attack. The Morning Download (Dutt, 2018; WSJ 2017; Norton, 2017) from 2017, with is analysis and adaptation to network behaviour is the second case. Allegedly, password and captcha attacks have also been observed in the wild (Norton, 2017). This report has identified use cases through the review of dual use technology research where explicit malicious perspective was absent. However, dual use is not the same as real malicious end use, nor is it definitive indicator or precursor to malicious end use. There are many known and unknown reasons to why certain types of malicious actors may or may not choose to implement AI in their cyberattack anatomy. The maturity of technology most likely matters, as well as accessibility, complexity of design and preparation, the existence of simpler or better alternatives, and notably the actual need for AI to achieve intended objectives.

7.3 Strategic capabilities for antagonists

The overall feature of AI is that it automates actions so that they become semi-automated or fully automated where they were previously manual. In some cases, AI can replace other forms of automation to facilitate more sophisticated actions. However, it should be noted that AI and other forms of automation may impact strategic capabilities in similar ways. Considering technologies that both exhibit a high level of readiness and proven antagonistic end use, there seems to be certain strategic capabilities facilitated by AI.

Aggregation: The use of AI to collect, collate, synthesize and analyze data faster than humanly possible will affect the speed, scale and precision of cyberattacks. CyberLover is an early demonstration of this impact with respect to precision, and the increasing number of AI threat hunting solutions on the market demonstrate the importance of scale in data processing on the cyber domain.

Repetition: The ability to generate more consistent and persistent repetition of tasks and actions than human attackers will affect the scale and magnitude of cyberattacks. CyberLover and captcha attacks as an early form of AI-supported cyberattacks, indicates this. However the current use of simpler tools, such as for password attacks (WSJ, 2017), indicate that this strategic capability is not particular for AI-automation.

Deception: The recent developments in GAN and other adversarial examples facilitates more effect social engineering campaigns, as demonstrated by SNAP_R. However, antagonistic use cases are not only about deception in AI versus human technology user scenarios as in CyberLover, but also AI versus technology, and even AI versus AI, as in the Morning Download and as demonstrated by Lee and Yim (2020), Chakraborty et al (2018), and Gu, Golan-Davitt and Garg (2019).

Manipulation: Given that AI identifies and operates on data and protocols faster, more consistently and with greater persistence than humans, its shows particular promise as a tool to exploit vulnerabilities in classifiers (e.g. spam filters) and similar technologies. This premise is supported by the findings on machine learning (ML) captcha attacks, research findings on classifier attacks, and research on natural language processing (NLP) attacks.

7.4 Defender – attacker arms race dynamics

The literature identified in this report demonstrates aspects of a defender-attacker artificial intelligence arms race. However, it cannot be decisively concluded whether such an arms race currently benefits defenders or attackers. Therefore, it is also not possible to conclude whether AI-supported cyber security or network defense solutions will be a sufficient countermeasure to artificially intelligent cyberattacks. On one hand, most of the research

found through the literature review was security-focused and had been carried out for the development of defense. Similarly, the report has identified several AI-supported cyber security tools on the market. In this respect, AI research and development seems to benefit defenders. On the other hand, both real cases of AI-augmented cyberattacks, as well as experiments conducted in the identified research indicate that AI cyber security solutions may not be enough to stave off AI-supported attackers. Classifier and NLP attacks are both AI-versus-AI examples where attackers can effectively foil intelligent defenses. Similarly, several identified solutions on the market indicate the defensive use of AI to detect abnormal and malicious activities within a network. However, with attackers using similar technologies (possibly combined with smart lateral movements, self-learning malware, discovery obfuscation, or “low-and-slow” exfiltration), it is not proven that such systems can learn to identify the next generation of intelligent attacks that blend in the network. A specific research case that deserves mention in this regard is Lee and Yim (2020) who both designed the AI security measure and the AI attack tool to foil it in their experiment concerning password stealing. Chakraborty et al, 2018, provide another excellent example of research on AI-versus-AI attacks.

8 Conclusions and further research

This report has accounted for an overarching literature review on the possible applications of artificial intelligence (AI) within the anatomy of cyberattacks. The review covered 96 publications and found 19 use cases. It found that the use cases that exhibit the highest technology readiness level primarily support the early phases of the cyberattack anatomy; reconnaissance in particular, but also access and penetration. Moreover, the report finds that real AI-supported cyberattacks that have occurred typically have the reconnaissance phase, access and penetration phase or the internal reconnaissance and lateral movement phases augmented with AI. The use cases presented in the results of this report should be regarded as a snapshot of currently known areas of research, rather than a prediction for future malicious end use. Other methods of research such as technical prognoses (Gisslén, 2014) and horizon scans (Karasalo & Schubert, 2019) can supplement this research with projects from different time perspectives. An approach supported by multiple research methods may be able to draw more concrete projections about notable advances in the sophistication of antagonistic capabilities.

The nature of the literature review results and their concrete findings give rise to several questions regarding the relationship between technical research and malicious end use. The security-focus and defense perspective necessitates further study, not only into the dual use nature of AI technology in this domain, but as a matter of substantive evaluation of potential spillover from security research into future malicious use. Which AI-tools will most likely be accessible to the future antagonist? Which categories of antagonists specifically might have access to such tools? Where will the future antagonist instead choose less advanced, cheaper, more effective or other more likely alternatives to AI? In answering, these questions, it may be necessary to conduct further research to identify open source “hacker assets” (Samtani et al, 2017) that can either be augmented with AI, fully automated through AI applications, or utilized in artificially intelligent cyberattacks. Additionally, at a state actor level, what policy initiatives might drive or intensify the development and applications of offensive capabilities in the future? Research aimed at capturing these aspects of technology development is currently being conducted and needs to be merged with technical analysis (e.g. Horowitz et al, 2018; Nato Hybrid Centre of Excellence 2019a & 2019 b).

A related question is how security-concerned researchers can be supported in predicting and avoiding future malicious end-use. In related contemporary debates about GAN and deep fake technology, a noteworthy concern has been raised within the AI community. The premise of this concern is that the bright and presumably well-intentioned minds of the field, trained in advanced research, may develop tools within a scientific culture of free exploration and open source publication, that are then adopted by less well-intentioned (and even less advanced) malicious actors (OpenAI, 2019).

An early ambition during the design of this study was to breach the gap between technical research and policy research, identifying tactics and strategies to counter artificially intelligent cyberattacks. However, the technical literature reviewed did not present concrete suggestions in this respect. It was therefore decided that this objective would be left out of the results. It should nevertheless be noted that the policy research collected in early phases of the study suggested predominantly hybrid warfare countering strategies (Nato Hybrid Centre of Excellence (2019a and 2019 b) or legal restrictions (UNIDIR, 2017a & 2017b) to AI development. For example, much of the legal debate has centered on human control of autonomous systems under international humanitarian law (UNIDIR, 2017a & 2017b; Guarino, 2013), and to some degree the criminal law aspects of artificially intelligent systems (e.g. Hallevy, 2015, 2019). On the basis of related research on the regulation of technology, we suggest that additional regulatory strategies exist but are currently not subject to extensive research; e.g. principles of harm avoidance and regulated ethics for technology developers (Resolution 2015/2103(INL)), export control and sanctions regimes (Zouave, 2017; Zouave & Vogiatzoglou, 2017), the application of criminal law and criminal deterrence strategies to malicious production and end use. Moreover, research on the

malicious uses of artificial intelligence in the digital domain underpins the need to begin identifying tools, techniques, methods and processes to operationalize strategies in to operational tactics to counter malicious development, proliferation and end use.

It is also noteworthy that given the deficit in research of countering tactics, the organizational resources needed for defense against artificially intelligent cyberattacks are not well understood. What type of resource acquisition and development will competent authorities need to effectively counter the future antagonist? What type of organizational processes and factor would strengthen defense in a potential AI arms race between defenders and antagonists?

9 References

Advanced Network Systems. (2018). *Anatomy of a Cyberattack.* Available: https://www.getadvanced.net/blog/article/anatomy-of-a-cyberattack. Last accessed 07/03/2019

Ahmed, H., Glasgow, J. (2012). Swarm Intelligence: Concepts, Models and Applications. (Queens University School of Computing)

Almukaynizi, M., Nunes, E., Dharaiya, K., Senguttuvan, M., Shakarian, J., Shakarian, P. (2017). Proactive Identification of Exploits in the Wild Through Vulnerability Mentions Online. *IEEE*

Anderson, H. S., Woodbridge, J., & Filar, B. (2016). DeepDGA: Adversarially-Tuned Domain Generation and Detection. In *arXiv [cs.CR]*. arXiv. http://arxiv.org/abs/1610.01969

Avgerinos, T., Brumley, D., Davis, J., Goulden, R., Nighswander, T., Rebert, A., Williamson., N. (2018). 'The Mayhem Cyber Reasoning System', *IEEE Security Privacy*, 16(2), pp. 52–60

Breck. (2020). *Detecting Cybersecurity Threats to AI.* Available: https://breckinc.com/2019/01/03/detecting-cybersecurity-threats-to-ai/. Last accessed 04/03/2020

Bursztein, E., Aigrain, J., Moscicki, A., Mitchell, J. C. (2014). The End is Nigh: Generic Solving of Text-based CAPTCHAs. *8th Usenix workshop on Offensive Technologies WOOT '14*

Bursztein, E., Beauxis, R., Paskov, H., Perito, D., Fabry C., Mitchell, J. (2011). The Failure of Noise-Based Non-continuous Audio Captchas, *2011 IEEE Symposium on Security and Privacy*, Berkeley, CA, 2011, pp. 19-31

Cakir, B., & Dogdu, E. (2018). Malware classification using deep learning methods. *ACMSE '18*

Castiglione, A., De Prisco, R., De Santis, A., Fiore, U., & Palmieri, F. (2014). A botnet-based command and control approach relying on swarm intelligence. *Journal of Network and Computer Applications*, *38*, 22–33

Chung, K., Kalbarczyk, Z. T., & Iyer, R. K. (2019). Availability attacks on computing systems through alteration of environmental control: smart malware approach. *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, 1–12

Correa Bahnsen, A. C., Ivan Torroledo, Luis David Camacho and Sergio Villegas. (2018). DeepPhish: Simulating Malicious *AI. 2018 APWG Symposium on Electronic Crime Research (eCrime).*

Batt, S. (2019-10-17). *How Artificial Intelligence Will Shape the Future of Malware.* Available: https://www.makeuseof.com/tag/artificial-intelligence-future-malware/. Last accessed 07/02/2020

BBC News (2016) *"Mayhem" wins hacking challenge*. Available at: https://www.bbc.com/news/technology-36980307 Last accessed 12/02/2020.

Bellman, R. E. (1978). *An Introduction to Artificial Intelligence: Can Computers Think?* San Francisco: Boyd & Fraser

Beran, T., Björnham, O. Gustavi, T., Hagström, M., Johansson, A., Karlholm, J., Lindblom, J., Oskarsson, D., Sommestad, T., Stensbäck, N., Svenmarck, P., Svensson, M., Önehag, A. (2017). Försvarsnära tillämpningar av Artificiell Intelligens. FOI-D--0807--SE

Biggio, B., Fumera, G., Roli, F. (2014). Security Evaluation of Pattern Classifiers under Attack. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. 26(4*)

Bilal, M., Gani, A., Lali, M. I. U., Marjani, M., and Malik, N. (2019). Social Profiling: A Review, Taxonomy, and Challenges. *Cyberpsychology, Behavior, and Social Networking*. 22 (7), 433-450

Blum, W. (2017). *Neural fuzzing: applying DNN to software security testing*. Available: https://www.microsoft.com/en-us/research/blog/neural-fuzzing/. Last accessed 21/02/2020

Boyd, J. (1987). A discourse on winning and losing. Maxwell Air Force Base, AL: Air University Library Document No. M-U 43947

Brundage, M., Avin, S., Clask, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.* Available: https://maliciousaireport.com/. Last accessed 27/02/2019

Brynielsson, J., Franke, U., Tariq, M. A., Varga, S.. (2016). Using Cyber Defense Exercises to Obtain Additional Data for Attacker Profiling. *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*

Brynielsson, J., Nilsson, M., Schubert, J., Svenmarck, P. (2018). Artificiell intelligens för beslutsstöd i ledningssystem. FOI-R--4678--SE

Brynjolfsson E., and McAffee, A. (2017). *The Business of Artificial Intelligence.* Available: https://starlab-alliance.com/wp-content/uploads/2017/09/The-Business-of-Artificial-Intelligence.pdf. Last accessed 26/11/2019.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D. (2018). Adversarial Attacks and Defences: A Survey. *ArXiv*

Choi, Y., Choi, M.,Kim, M., Ha, J-W., Kim, S., Choo, J. (2018). StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *IEEE Xplore*

Danziger, M., & Henriques, M. A. A. (2017). Attacking and Defending with Intelligent Botnets. *XXXV SBrT*

DARPA. (2015). *Seven Teams Hack Their Way to the 2016 DARPA Cyber Grand Challenge Final Competition.* Available: https://www.darpa.mil/news-events/2015-07-08. Last accessed 27/02/2019

DARPA (2016) *CGC Results.* Available: http://archive.darpa.mil/cybergrandchallenge/event.html Last accessed 01/10/2018

DARPA. (2017). *Cyber Grand Challenge (CGC) (Archived).* Available: https://www.darpa.mil/program/cyber-grand-challenge. Last accessed 27/02/2019

De Gregorio, A. (2016) Vulnerabilities and Their Surrounding Ethical Questions: A Code of Ethics for the Private Sector. *IEEE International Conference on Cyber Conflict, 21-23 October, Washington, DC, USA*

Dhaya, R., and Poongodi, M. (2014). Detecting software vulnerabilities in android using static analysis.

Dheap, V. (2017). *AI in Cybersecurity: A Balancing Force or a Disruptor?* Available: https://www.rsaconference.com/industry-topics/presentation/ai-in-cybersecurity-a-balancing-force-or-a-disruptor. Last accessed 13/02/2020

Darktrace. (2018). *The Next Paradigm Shift: AI-Driven Cyber-Attacks.* Available: https://www.darktrace.com/en/resources/wp-ai-driven-cyber-attacks.pdf. Last accessed 27/02/2019

DeepMind. (2019). *AlphaStar: Grandmaster level in StarCraft II using multi-agent reinforcement learning*. Available: https://deepmind.com/blog/article/AlphaStar-Grandmaster-level-in-StarCraft-II-using-multi-agent-reinforcement-learning. Last accessed: 20/02/2020

Dixon, W., Eagan, N. (2019-06-19). *3 ways AI will change the nature of cyber attacks*. Available: https://www.weforum.org/agenda/2019/06/ai-is-powering-a-new-generation-of-cyberattack-its-also-our-best-defence/. Last accessed 03/02/2020

Dutt, D. (2018). *2018: the year of the AI-powered cyberattack.* Available: https://www.csoonline.com/article/3246196/2018-the-year-of-the-ai-powered-cyberattack.html. Last accessed 02/03/2020.

European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))

Fahad, A., Yahya, S. K. Abdulsamad, E. (2018). Inflectional Review of Deep Learning ozwxsen Natural Language Processing. *International Conference on Smart Computing and Electronic Enterprise. (ICSCEE2018)*

Falco, G., Viswanathan, A., Caldera, C., and Shrobe, H. (2018). A Master Attack Methodology for an AI-Based Automated Attack Planner for Smart Cities. *IEEE Access 6,* pp. 48360-48373

Al Fahdi, M., Clarke, N. L., Furnell, S. M. (2013). Towards An Automated Forensic Examiner (ARE) Based Upon Criminal Profiling & Artificial Intelligence. *Australian Digital Forensics Conference 2013*

Filippoupolitis, A., Loukas, G., Kapetanakis, G. (2014). Towards real-time profiling of human attackers and bot detection. *CFET 2014 7th International Conference on Cybercrime Forensics Education & Training*

Flowmon. (2020). *Network Behavior Analysis & Anomaly Detection*. Available: https://www.flowmon.com/en/solutions/security-operations/network-behavior-analysis-anomaly-detection. Last accessed 02/03/2020

Fruhlinger, J. (2017) *Petya ransomware and NotPetya malware: what you need to know*. Available: https://www.csoonline.com/article/3233210/petya-ransomware-and-notpetya-malware-what-you-need-to-know-now.html. Last accessed 03/02/2020

Ghallab, M., Nau, Dana S.; Traverso, P. (2004). Automated Planning: Theory and Practice, Morgan Kaufmann, ISBN 1-55860-856-7

Gisslén, L. (2014). Artificiell intelligens: Teknisk prognos. FOI-R--3919--SE

Goertzel, B. and Pennachin, C. eds (2007). *Artificial General Intelligence*. Heidelberg: Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press

Greeff, E., Ross, W. (2019). The Rise of Machines, AI- and ML-Based Attacks Demonstrated. *RSA Conference*. Available: https://www.conferencecast.tv/talk-16913-the-rise-of-the-machines-ai-and-ml-based-attacks-demonstrated. Last accessed: 25/02/2020

Guarino, A. (2013). Autonomous Intelligent Agents in Cyber Offence. *2013 5th International Conference on Cyber Conflict*

Gu, T., Golan-Davitt, B., Garg, S. (2019). BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *IEEE*

Hagström, M. (2016). Autonom våldsutövning – hot eller möjligheter: För en strukturerad debatt. FOI Memo 5676

Hallevy, G. (2015). *Liability for Crimes Involving Artificial Intelligence Systems*. New York: Springer

Hallevy, G. (2015). The Basic Models of Criminal Liability of AI Systems and Outer Circles. *SSRN*

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge: MIT Press

Herr, T. (2014). PrEP: A Framework for Malware & Cyber Weapons. *The Journal of Information Warfare*. 13 (1)

Hitaj, B., Gasti, P., Ateniese G., Perez-Cruz, F. (2018). PassGAN: A Deep Learning Approach for Password Guessing. *NeurIPS 2018 Workshop on Security in Machine Learning (SecML'18)*

Harvey, P. (2019). *ExifTool by Phil Harvey: Read, Write and Edit Meta Information!* Available: https://www.sno.phy.queensu.ca/~phil/exiftool/. Last accessed 26/11/2019.

Holm, H., and Sommestad, T. (2016). SVED: Scanning, Vulnerabilities, Exploits and Detection. *Cyber Security and Trusted Computing, Milcom 2016, track 3*.

Horowitz M., Scharre, P., Allen, G.C., Frederick, K., Cho, A., Saravalle, E. (2018). *National Security-Related Applications of Artificial Intelligence.* Available: https://www.cnas.org/publications/reports/artificial-intelligence-and-international-security. Last accessed 27/02/2019

Hunters.AI. (2020). *Autonomous Threat Hunting*. Available: https://hunters.ai/. Last accessed 02/03/2020.

Hybrid CoE. (2019a). *Hybrid Warfare – Orchestrating the Technology Revolution*. Available: https://www.hybridcoe.fi/news/hybrid-warfare-orchestrating-the-technology-revolution/. Last accessed 27/02/2020

Hybrid CoE. (2019b). *Hybrid Warfare and New Technologies workshop in Stockholm*. Available: https://www.hybridcoe.fi/news/hybrid-warfare-workshop-in-stockholm/. Last accessed 27/02/2020

ImmuniWeb. (2020). We Reduce Complexity and Costs of Application Security. Available: https://www.immuniweb.com/. Last accessed 02/03/2020

Indurkhya, N., Damerau, F. J. (2010). Handbook of Natural Language Processing. Coca Raton: Taylor G Francis.

Jones, C. L., Bridges, R. A., Huffer, K. M. T., Goodall J. R. (2015). Towards a Relation Extraction Framework for cyber-security concepts. *Cyber and Information Security Research Conference 2015*. 1-4

Juric, R., Moholth McClenaghan, K., Moholth, O. C. (2019). Detecting Cyber Security Vulnerabilities through Reactive Programming. *Proceedings of the 52nd Hawaii International Conference on System Science 2019*. 7204-7213

Kali Tools. (2019). *theharvester Package Description.* Available: https://tools.kali.org/information-gathering/theharvester. Last accessed 27/11/2019

Kapetanakis, S., Filippoupolitis, A., Loukas, G., Al Murayziq, T. S.. (2014). Profiling cyber attackers using Case-based Reasoning. UK-CBR 2014

Karasalo, M., Schubert, J. (2019). Developing Horizon Scanning Methods for the Discovery of Scientific Trends. *2019 International Conference on Document Analysis and Recognition (ICDAR).*

Kirat, D., Jang, J., Stoecklin, PH.M. (2018). *DeepLocker: Concealing Targeted Attacks with AI Locksmithing.* Available: https://i.blackhat.com/us-18/Thu-August-9/us-18-Kirat-DeepLocker-Concealing-Targeted-Attacks-with-AI-Locksmithing.pdf. Last accessed 27/02/2019

Kubovič, O., Košinár, P., & Jánošík, J. (2018). *Can artificial intelligence power future malware?* Available: https://www.welivesecurity.com/wpcontent/uploads/2018/08/Can_AI_Power_Future_Malware.pdf. Last accessed 18/03/2020

Lando, G. (2018-08-27). *Machine Vs. Machine. A Look at AI-powered Ransomware.* Available: https://www.getfilecloud.com/blog/2018/08/machine-vs-machine-a-look-at-ai-powered-ransomware/#.XjvenTFKg2x . Last accessed 04/02/2020

Lee, K., and Yim, K., (2020). Cybersecurity Threats Based on Machine Learning-Based Offensive Technique for Password Authentication. *Applied Science 10(4)*

Lima, F. M, Sampaio, L.D.H, Rodrigues J, Zarpelao, B.B (2010). Anomaly detection using baseline and K-means clustering. *Conference Paper. International Conference on Software, Telecommunications and Computer Networks*

Lin, Y-C., & Hong, Z-W., Liao, Y-H., Shih, Meng-Li, S., & Liu, M-Y., & Sun, M. (2017). Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. *IJCAI 2017*. 3756-3762. 10.24963/ijcai.2017/525

Lockheed Martin. (2015). *Gaining the Advantage: Applying Cyber Kill Chain Methodology to Network Defense.* Available: https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/Gaining_the_Advantage_Cyber_Kill_Chain.pdf. Last accessed 19/02/2020.

Luckow, K., Kersten, R., and Pasareanu C. (2020). Complexity vulnerability analysis using symbolic execution. *Journal of Software: Testing, Verification and Reliability*

Ludwig, S. (2017). Intrusion detection of multiple attack classes using a deep neural net ensemble. *Conference: 2017 IEEE Symposium Series on Computational Intelligence (SSCI).* 1-7. 10.1109/SSCI.2017.8280825.

Löfvenberg, J., Sommestad, T., and Bildsten, C. (2019). Automatisk attackkodsgenerering: En skanning av forskningsfronten. FOI-R--4737--SE

Ma, W., Duan, P., Liu, S., Gu, G., Liu, J. (2012). Shadow attacks: automatically evading system-call-behavior based malware detection. *Journal in Computer Virology 8(1-2),* 1–13. https://doi.org/10.1007/s11416-011-0157-

McCarthy, J. (2007). *Basic Questions.* Available: http://www-formal.stanford.edu/jmc/whatisai/node1.html. Last accessed 26/03/2019.

McDermott, D. (1985). *Introduction to Artificial Intelligence*. Boston: Addison Wesley

McMillan, R. (2010). *Wiseguy scalpers bought tickets with CAPTCHA-busting botnet.* Available: https://www.computerworld.com/article/2514577/wiseguy-scalpers-bought-tickets-with-captcha-busting-botnet.html. Last accessed 04/03/2020

Metivier, B. (2018). *Threat Hunting: Anatomy of a Cyber Attack.* Available: https://www.sagedatasecurity.com/blog/threat-hunting-anatomy-of-a-cyber-attack. Last accessed 07/03/2019

Miktkov, R (2003). The Oxford Handbook of Computational Linguistics. New York: Oxford University Press

MITRE. (2015). *JUST RELEASED: ATT&CK for Industrial Control Systems.* Available: https://attack.mitre.org/. Last accessed 18/03/2020

Moisejevs, I. (2019). *Evasion attacks on Machine Learning (or "Adversarial Examples").* Available: https://towardsdatascience.com/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1?gi=2daefa4d5f32. Last accessed 06/03/2020

Mulwad, V., Li, W., Joshi, A., Finin, T., and Viswanathan, K. (2011). Extracting Information about Security Vulnerabilities from Web Text. *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. 257-260

Murphy, K. P. (2010). *Machine Learning: A Probabilistic Perspective.* Cambridge: The MIT Press

Nadler, A. (2019). *Adversarial DGA - Is It Out There?* Available: https://blogs.akamai.com/sitr/2019/07/adversarial-dga---is-it-out-there.html. Last accessed 05/03/2020

NASA. (2012). *Technology Readiness Level.* Available: https://www.nasa.gov/directorates/heo/scan/engineering/technology/txt_accordion1.html. Last accessed 05/02/2020

Norton, S. (2017). *Era of AI-Powered Cyberattacks Has Started.* Available: https://blogs.wsj.com/cio/2017/11/15/artificial-intelligence-transforms-hacker-arsenal/. Last accessed 27/02/2019

OpenAI. (2019). *Better Language Models and Their Implications.* Available: https://openai.com/blog/better-language-models/. Last accessed 05/03/2020

Oracle. (2017). *Anatomy of a Cyber Attack: The Lifecycle of a Security Breach.* Available: http://www.oracle.com/us/technologies/linux/anatomy-of-cyber-attacks-wp-4124673.pdf. Last accessed 07/03/2019

Perera, I., Hwang, J., Bayas, K., Dorr, B., Wilks,Y. (2018). Cyberattack Prediction Through Public Text Analysis and Mini-Theories. *2018 IEEE International Conference on Big Data (Big Data)*. 3001-3010

Pfleeger, S. (2010). Anatomy of an Intrusion. *IT Professional*. 12 (4), 20-28

Randhawa, S., Turnbull, B., Yuen, J, Dean, J. (2018) Mission-centric Automated Cyber Red Teaming. ARES 2018 1-11.

Rhode, B. (ed.). (2019) Artificial intelligence and offensive cyber weapon. *Strategic Comments 25(40).* DOI: 10.1080/13567888.2019.1708069

Rice, A. (2014). *Command-and-control servers: The puppet masters that govern malware*. Avaialble: https://searchsecurity.techtarget.com/feature/Command-and-control-servers-The-puppet-masters-that-govern-malware. Last accessed: 18/03/2020

Rosell, M., Bolin, U., Brynielsson, J., Garcia Lozano, M., Gustafsson, D., Horndahl, A., Karasalo, M., Lilja, H., Pelzer, B., Stenborg, K-G., Valldro, E., Varga, S. (2018). Semi-automatisk datadriven webbanalys: forskning, prototyputveckling och undersökningar. FO-R--4692--SE

Rossi, S. (2007). *Beware the CyberLover that Steals Personal Data.* Available: https://www.pcworld.com/article/140507/article.html. Last accessed 19/02/2020.

Samtani, s., Chinn, R., Chen, H., and Nunamaker J. F. Jr. (2017). Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence. *Journal of Management Information Systems*. 34 (4), 1023-1053.

San Agustin, C. E. (2019). Mitigating Deep Learning Vulnerabilities from Adversarial Examples Attack in the Cybersecurity Domain. *CES Agustin. arXiv preprint*

Scharre, P., and Horowitz, M. (2018). *Artificial Intelligence What Every Policymaker Needs to Know.* Available: https://www.cnas.org/publications/reports/artificial-intelligence-what-every-policymaker-needs-to-know. Last accessed 27/02/2019

Schroer, A. (2020). *30 companies merging AI and cybersecurity to keep us safe and sound*. Available: https://builtin.com/artificial-intelligence/artificial-intelligence-cybersecurity. Last accessed 02/03/2020

Schubert, J. (2017). Artificiell Intelligens för Militärt Beslutsstöd. FOI-R--4552--SE

Seymour, J. and Tully, P. (2016). *Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter.* Available: https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf. Last accessed 06/02/2020

Seymour, J. and Tully, P. (2017). Generative Models for Spear Phishing Posts on Social Media. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*

Siegel, M. (2003). The Sense-Think-Act Paradigm Revisited (International Workshop on Robotic Sensing: Örebro, Sweden, 5-6 June 2003)

Sood, A. K., & Zeadally, S. (2016). A Taxonomy of Domain-Generation Algorithms. *IEEE Security Privacy, 14(4).* 46–53

Sovereign Intelligence. (2018). *AI-Driven Intelligence*. Available: https://www.sovereign.ai/. Last accessed 02/03/2020.

Spiderfoot. (2019). *About.* Available: https://www.spiderfoot.net/documentation/#spiderfoot-hx. Last accessed 27/11/2019

Stapleton, K., and Stevens, Y. (2019). *AI Powered Malware: The New Frontier for Cybersecurity*. Available: https://medium.com/@ystvns/ai-powered-malware-the-new-frontier-for-cybersecurity-3520dce3a138 Last accessed: 18/03/2020

Stoecklin, M. (2018). *DeepLocker: How AI Can Power a Stealthy New Breed of Malware.* Available: https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/. Last accessed 18/03/2020

Strömberg, D. (1987). Anteckningar om AI. Del 1: Teknisk prognos. FOA C 30448-3.3

Suganya, G., Karpgavalli, S., and Christina, V., (2010). Proactive Password Strength Analyzer Using Filters and Machine Learning Techniques. *International Journal of Computer Applications 7(14)*

Svenmarck, P., Wikström, M., Zouave, E., Krona, M. (2020). Artificiell intelligens för obemannade luftfartyg inom krishantering: Möjligheter och hot. (Forthcoming)

Svensson E., Magnusson, J., Zouave E. (2019). Kryptomaskar och deras konsekvenser: Åtgärder för cybersäkerhet utifrån fallen WannaCry och NotPetya. FOI-R--4774--SE

Swedish Government Offices. (2016). Nationell strategi för samhällets informations- och cybersäkerhet. Skr. 2016/17:213

Szmit, M., Szmit, A. (2012). Usage of Modified Holt-Winters Method in the Anomaly Detection of Network Traffic: *Case Studies. Journal of Computer Networks and Communications 2012*

Teahan, J. W. (2010). *Artificial Intelligence – Agents and Environments*. Telluride: Ventus

Trieu, K., and Yang, Y. (2018). Artificial Intelligence-Based Password Brute Force Attacks. *MWAIS 2018 Proceedings*

Truong, C. T. Zelinka, I., Plucar, J., Candik, M., Sulz, V. (2020). ”Artificial Intelligence and Cybersecurity: Past, Presence, and Future” in. (2020). Artificial Intelligence and Evolutionary Computations in Engineering Systems by Lakshmi, C., Das, S., Panigrahi, B. 10.1007/978-981-15-0199-9.

UNIDIR. (2017a). *The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches* (No. 6).Available: http://www.unidir.org/files/publications/pdfs/the-weaponization-of-increasingly-autonomous-technologies-concerns-characteristics-and-definitional-approaches-en-689.pdf. Last accessed 27/02/2019

UNIDIR. (2017b). The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations. Report no. 7 (2017)

Vectra. (2019). *The Cognito platform*. Available: https://www.vectra.ai/product/what-it-is. Last accessed 02/03/2020

Vijaya, M. S., Jamuna, K. S., and Karpagavalli, S. (2009). Password Strength Prediction Using Supervised Machine Learning Techniques. *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*

Waddell, K. (2016). *The Twitter Bot That Sounds Just Like Me.* Available: https://www.theatlantic.com/technology/archive/2016/08/the-twitter-bot-that-sounds-just-like-me/496340/. Last accessed 2020-02-06

Wang, Y., Huang, Y., Zheng, W., Zhou, Z., Liu, D., Lu, M. (2017). Combining convolutional neural network and self-adaptive algorithm to defeat synthetic multi-digit text-based CAPTCHA. *2017 IEEE International Conference on Industrial Technology (ICIT)*

Wang, Z., Zhang, Y., Tian, Z., Ruan, Q., Liu, T., Wang, H., Liu, Z., Lin, J., Fang, B., Shi, W. (2019). Automated Vulnerability Discovery and Exploitation in the Internet of Things. *Sensors*

Wirkuttis, N., Klein, H. . (2017). Artificial Intelligence in Cybersecurity. *Cyber, Intelligence, and Security*. 1 (1), 103-119.

WSJ. (2017). *The Morning Download: First AI-Powered Cyberattacks Are Detected.* Available: https://blogs.wsj.com/cio/2017/11/16/the-morning-download-first-ai-powered-cyberattacks-are-detected/. Last accessed 02/03/2020

Yuen, J., Turnbull, B., and Hernandez, J. (2015). Visual analytics for cyber red teaming. *2015 IEEE Symposium on Visualization for Cyber Security (VizSec), Chicago, IL,* pp. 1-8

Yuen, J. (2013). Automated Cyber Red Teaming. Cyber and Electronic Warfare Division in Defence Science and Technology Organisation. DSTO-TN-1420

Zetter, K. (2010). *Wiseguys Plead Guilty in Ticketmaster Captcha Case.* Available: https://www.wired.com/2010/11/wiseguys-plead-guilty/. Last accessed 04/03/2020

Zhang, W. E., Sheng, Q., & Alhazmi, F., Li, C. (2019). Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey 1(1). Association for Computing Machinery

Zhang, S., Ou, X., and Carragea, D. (2015). Predicting Cyber Risks through National Vulnerability Database. *Information Security Journal: A Global Perspective 24(4-6)*

Zhou, A. (2018). Bringing the Fight to Them: Exploring Aggressive Countermeasures to Phishing and other Social Engineering Scams. (Comp116 Final Paper)

Zouave, E., Vogiatzoglou, P. (2017). *Dual Use Technology Controls for Security Researchers Of Double-Edged Swords & Blunt Bludgeons.* Available: https://www.law.kuleuven.be/citip/blog/dual-use-technology-controls-for-security-researchers-of-double-edged-swords-blunt-bludgeons/. Last accessed 05/03/2020

Zouave, E. (2017). *Researcher Compliance with Export Controls – Blunting the Sword & Honing the Bludgeon.* Available: https://www.law.kuleuven.be/citip/blog/researcher-compliance-with-export-controls-blunting-the-sword-honing-the-bludgeon/. Last accessed 05/03/2020

Zouave, E. (2019). Aktiva operationer på cyberdomänen: Folkrättslig normativ utveckling. FOI-R--4776--SE

FOI

ISSN 1650-1942

www.foi.se