WEAPONSSECURITY COMPLY OR EXPLAIN

(un)certainties in the cybersecurity of the weaponization of autonomy in systems

Leiden University Master thesis for the Executive Master Cyber Security
Student: R.R.J. van de Pol

First supervisor: Prof. dr. B. van den Berg, Leiden University
Second supervisor: Prof. dr. F.P.B. Osinga, Leiden University

February 2021

Abstract

Digitalization is undoubtedly part of the military domain. The discussion about the cybersecurity of autonomous weapon systems lacks a debate. Since November 2019, the Group of Governmental Experts (GGE) on emerging technologies (United Nations) in the area of lethal autonomous weapons agreed on a set of non-binding guiding principles. Principle (f) states: "When developing or acquiring new weapons systems based on emerging technologies in the area of lethal autonomous weapons systems, physical security, appropriate non-physical safeguards (including cyber-security against hacking or data spoofing), the risk of acquisition by terrorist groups and the risk of proliferation should be considered [1, p. 10]" But what are the defensive cybersecurity risks for (autonomous) weapon systems, and how can regulatory intervention help increase such systems' defensive cybersecurity? The discussion about autonomy as a general attribute is imprecise. One of the major challenges is the conflicting descriptions of autonomy from either a general sense, or in reference to a particular context or system. The definitions lack a system-centric approach, adding the element of human-machine and thus the possibility of meaning full human control defines the autonomy in the system especially in the targeting function of the weapon itself. The definition of autonomous weapon systems can be more specific if it entails also the political, legal and military decision making process that takes place on before, during and after the actual occupation of the weapon since the risk of an unintended casualty potentially can start there. The identification of basic information security risks showed that standards and frameworks for securing exists. But the detailed engineering operations of the protocols, the never ending process of identifying critical vulnerabilities, and understanding how to address these vulnerabilities without compromising functionality, are so complex that is hard to make a trade-off in order to achieve functionality by accepting a certain level of vulnerabilities. The ethical discussion around the possible lack of transparency, combined with the risks of data bias and the need to have full control over data makes it important to develop requirements for making algorithms secure, reliable and robust with a trusted community within the ecosystem. Via standard-setting and norms on multilateral level and through public private cooperation, the market can be steered towards a secure development. Regulation mitigates the risks that nations will have an incentive to rapidly acquire and integrate systems without placing appropriate cybersecurity policies to ensure that systems are safe and reliable. Standard-setting can gain a strategic value for the weapon systems' cybersecurity. Therefore the strategy must include goals to, develop common definitions, create transparency and ensure compliance of the standards.

Content

1. Introduction: weapons in connection

Digitalization is undoubtedly part of the military domain. The need to adjust to and incorporate new information systems and innovation trough digitalization into Europe's armed forces is essential for defending society against threats. For decades, algorithms have helped the military identify targets, estimate harm, and launch direct attacks as a part of weapons systems [2]. Algorithms that learn to sense and help systems move through the battlefield can increase military power [3]. Civil trends likely will shape the military industry also over the next few years. Most prominent examples, such as Artificial Intelligence (A.I.), Internet of Things (IoT), and the upcoming mobile technology 5G, show a growing number of connected devices. These developments and new technologies have the potential to achieve severe cost-savings and optimize military operations. An in-depth study from the American Government Accountability Office in 2018 [4] concludes that weapon systems are more software-dependent and networked than ever before, and automation and connectivity are fundamental enablers of modern military capabilities.

Due to digitalization, military conflicts are shifting from the physical world to the virtual world, also known as cyberspace [5]. Military conflicts no longer only appear to be on land, in the air, at sea, but also in the digital domain [6], [7] or via a hybrid mix of all those domains. Furthermore, scholars even speak of an ongoing international arms race [8]. Reports show that Russia is field testing several Artificial Intelligence-enhanced unmanned-systems in Syria [8]. In the U.S. alone, the Department of Defense possesses nearly 11,000 unmanned aerial systems of many different types and capabilities that function via Artificial Intelligence [9].

Simultaneously, there are also adverse effects: digitization can result in new vulnerabilities leading to insecurity or disorder in the civilian and military domain. The current scientific debate mostly tries to understand risks that could potentially have a harmful impact, which leads to discussions about fundamental concerns like ethics, trustworthiness and responsible use [10]–[12]. Ethical scholars express severe reservations about the legal and ethical implications of using Lethal Autonomous Weapon System (LAWS). These reservations are concerns of an artificial intelligence-induced apocalypse regarding human rights expectations, and these scholars fear insecurity in an automated world [13]–[16]. They project the unknown results and the unpredictability that technological change can have into the debate about practical implications to human affairs. Yampolskiy, Spellchecker and Pistono [17]–[19] argue that autonomous systems may surpass humans and that humans thus may lose control. Alternatively, these scholars fear that an autonomous system might run amok [20]. Their concerns include the way algorithms seem to be organizing human life outside of our direct control.

Their concerns are that society has become inhabited by algorithms or code operating mostly implicitly, placing human control in the background [14].

Nevertheless, critics argue that these visions are somewhat divorced from the war's realities and how existing weapons use forms of autonomy [21]. As Birkeland [22] argues, war is inherently a human activity, and humans are responsible for a military mission, even if an autonomous system executes that mission. With different levels of autonomy come varying levels of control [22]. However, when it comes to decisions of life or death, the control over weapon systems must necessarily be meaningful in order to give legitimacy to the delegation of decision-making from the human to the machine. Currently, most weapon systems are —to some extent—controlled by a human operator [23]. Technological advancements, such as Artificial Intelligence or earlier the use of an Israeli uninhabited combat aerial vehicle (HARPY) [23], have sparked ethical, legal and moral discussions about the appropriateness of this innovations.

On a strategic level, risks will follow the fast introduction of new technologies, such as Artificial Intelligence (A.I.) in autonomous weapon systems, that might increase the likelihood and risk of war, leading to the escalation of ongoing conflicts, which will proliferate to malicious actors. This transformation can change the conduct of war [24]. Moreover, it has sparked ethical, legal and moral discussions about the appropriateness of these innovations. Increasing autonomy in weapon systems also raises a complicated set of accountability concerns [25], questions concerning applicable legal regimes [26], responsibility [27], moral codes and cultural beliefs [28], [29]. While significant legal concerns have been charged against fully autonomous offensive weapons, this is not further discussed in detail, since that is not the purpose of this thesis.

Impressive videos on Youtube of robots and drones give the impression that those machines function autonomously. However, in practice today, autonomous systems only appear to operate in a controlled environment [21], [30], [31]. According to Boulanin and Verbruggen [32], the discussion about autonomy as a general attribute is imprecise. In their view autonomy "may serve different capabilities in different weapon systems, and the concerns should – be they legal, ethical or operational – be articulated on the use of specific tasks; therefore this discussion will benefit from a more platform or system-centric approach" [32, p. 18].

A survey conducted by Jahan et al. [33] shows that the discussion about the cybersecurity of autonomous weapon systems (AWS) lacks a debate. Jahan et al. conclude that autonomous systems' cybersecurity is at a very early stage. However, maybe cybersecurity can improve military readiness through the optimization and relevance of weapons systems. Weapon systems are, in that sense, just like information technology (I.T.) or operational technology (O.T.), vulnerable to cyberattacks. Even though standard security measures, such as air-gapping, ensure that a weapon system lacks a direct connection to other untrusted systems or networks, attackers might exploit external interfaces (e.g., radios, radars, and maintenance ports) [34]. Hackers might try to gain access to weapon systems' internal computers, networks, and data [34], or when weapons-systems are using new technologies such as A.I., the system runs the risk of being manipulated or misled, which may result in security implications [35], [36]. Alternatively, McGraw [37] claims that: "cyberwar, cyber-espionage, and cybercrime all share the same root cause: our dependence on insecure networked computer systems [37, p. 111]. "Therefore defending against a cyberattack (by building security in) is just as important as developing offensive measures [37, p. 113]."

Technology has conceivably led to new attack methods using Advanced Persistent Threats (APT) or a cyberweapon. According to Rid: "cyberweapons are (computer)code that is used- or designed to be used- to threaten or cause physical, functional, or mental harm to structures, systems, or living beings [6, p. 7]. An example of a cyberweapon is the malware (malicious code) known as Stuxnet [38], [39]. A possible scenario from a state-actor could be a mix of conventional and cyber effects, such as malicious code combined with an intelligence operation to plant the code [39]. Operational risks, such as those related to reliability [40], fragility [41], and security of systems [42], [43] could raise fundamental questions about whether military systems will function and operate according to military commanders' intent [20]. The threat that malicious users hack or control such systems by exploiting their vulnerabilities could be realistic [33]. Since these autonomous systems are still evolving, it is essential to analyze these machines' cybersecurity issues. The state actor scenario requires appropriate defence strategies of cybersecurity, which might make the concerns mentioned above, such as the protection of unintended use, even more critical.

1.1. The lack of cybersecurity in legacy weapons systems

Innovative strategies such as Network-centric Warfare [44]–[47] during the first decade of the 20th century tried to achieve connections and communication between weapon systems and Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance systems (C4ISR systems) [46]. Connecting new and older systems helped achieve tactical means to increase the situational awareness needed for Battlefield Management Systems (BMS) [48]. Better situational awareness may help commanders to make better decisions.

Older weapon systems are not necessarily designed with cybersecurity in mind [45]. To avoid the risk that attackers can exploit their initial access to disrupt or degrade a weapon system's operation, updating the system is necessary. The authors of a report of the American Government Accountability Office [4] wrote: "Due to this lack of focus on weapon systems cybersecurity, DOD likely has an entire generation of systems that were designed and built without adequately considering cybersecurity.

Bolting on cybersecurity late in the development cycle or after the system's deployment is more difficult and costly than taking security into account during the design [4, p. 18]

Technical issues are also a reason for more attention to cybersecurity. The algorithm's technological complexity makes the decision-making-logic challenging to understand and explain [49]. According to Lewis et al.[2], complexity explains the necessity to discern whether the use of an emerging technological capability would create unknown risks. The development of weapons systems might bring new forms of risk that need to merge with algorithmic systems into military functions [24].

1.2. A need for regulated cybersecurity of future weapon systems?

The use of digital technologies in weapon systems by militaries is growing; these technologies' consequences are mediated by military organizations' ability to use them effectively. For the state's legitimacy, citizens must trust the supervision and appropriate use of technology by the government and military. Security-policies are seen as the pursuit of protecting life and preventing uncertainties from happening; hence security is the search for removing these uncertainties [50]. Nevertheless, scholars [51], [52] claim that international competition in developing autonomous weapon systems could escalate into an arms race. Without sufficient attention on responsible and secure development of emerging new technologies like Artificial Intelligence, nations might encourage rapidly acquiring and integrating autonomous weapon systems without putting appropriate policies in place to ensure that systems are safe and reliable [41]. Only since November 2019, the Group of Governmental Experts (GGE) on emerging technologies (United Nations) in the area of lethal autonomous weapons agreed on a set of non-binding guiding principles [1]. The CCW High Contracting Parties' meeting decided to endorse 11 "guiding principles" affirmed by the GGE. Principle (f) states: "When developing or acquiring new weapons systems based on emerging technologies in the area of lethal autonomous weapons systems, physical security, appropriate non-physical safeguards (including cyber-security against hacking or data spoofing), the risk of acquisition by terrorist groups and the risk of proliferation should be considered [45]" Meaning that the GGE agreed on using the principles as a basis for the clarification, consideration and development of aspects of the normative and operational framework on emerging technologies in the area of lethal autonomous weapons systems. The intention of the CCW is to use these principles to develop a "normative and operational framework" with the goal of presenting this framework at the 2021 Sixth Review Conference.

However, these principles are a start that stresses the importance of cybersecurity. To apply this principle, a deeper understanding of what needs to be regulated is necessary. Therefore, a first step towards regulating is awareness of the heterogeneity of secure development norms for weapon systems. Even then, regulation remains a complex policy problem to shape. Weapon systems include a wide range of different technologies and are systems-of-systems [53]: a collection of systems, each capable of independent operation, that interoperate together. The past shows that critical technological developments, such as nuclear weapons, are heavily regulated. Regulation is a tool to support governance [54] and can apply to various forms of inquiry shaping social behaviour, state and non-state standard-setting, monitoring and behaviour modification processes [32]. Thus regulation does not always focus solely on the law. Law is only one of the instruments that policymakers can use. Different techniques and instruments are available for the problems that policymakers are trying to solve. Regulation of weapon systems' cybersecurity might also play a strategic role, as an instrument of a government to steer, not by waving a raised finger, but by providing frameworks to promote weapons systems' development responsibly. In this thesis, the following research question will be answered:

What are the defensive cybersecurity risks for (autonomous) weapon systems, and how can regulatory intervention help increase such systems' defensive cybersecurity?

This research question can be divided into the following sub-questions:

- *What are weapon systems, and how can autonomy in weapons be defined?*
- *Which cybersecurity risks can be defined for weapon systems?*
- *To what extent can regulatory strategies help to achieve secure weapon systems?*

1.3. Research methods

This research includes questions about the need for regulations targeted explicitly to regulate autonomous weapon systems by applying cybersecurity. It expands knowledge by investigating how regulation variants can help to stay in charge of weapon systems via cybersecurity functions within their systems.

This thesis tries to increase understanding and explain a complicated issue with multiple relations and interests based on a qualitative approach. From this view, it combines descriptive and prescriptive perspectives.

This thesis describes the exploratory research carried out for this purpose. It consists of collecting, analyzing and interpreting theories of regulation, international relations, military strategies, automated and autonomous (weapon) systems, and cybersecurity. This thesis primarily uses an inductive process for understanding the complexity of secure weapon systems. The emphasis is on evaluating, not on creating new designs or models. The focus is on perspective and relative importance.

This study is interpretative. The methodological approach taken in this study is a mixed methodology based on desk research and semi-structured interviews. For classification reasons, only publicly available sources are used. Specific literature is collected using the snowball-method. The problem is that a prescriptive policy might not be theoretical enough. Therefore a clear theoretical framework of risk identification and regulation is set up. The analysis will focus on how and which different options are possible. Subsequently, interview data is interpreted and analyzed, and as a final step, conclusions are drawn on the ideas presented in theory during the literature review. To learn more about the experts' experiences and perspectives in military organizations seven interviews are held: two with cybersecurity experts, two with technical cybersecurity experts, one with senior data specialist and two interviews with senior policy advisors. The semi-structured interview protocol is added in appendix I.

1.4. Structure of this thesis

This thesis proceeds in three chapters. The next chapter covers an overview of weapon systems, focussing on their current or future possibilities and necessity. There is a small outlook into the future, but it mostly describes weapon systems' current practical application. This chapter is based on the extensive report of the researchers Boulanin and Verbruggen [5]. They performed a one-year mapping on the development of autonomy in weapon systems and used an extensive data set of weapon systems.

The third chapter investigates the cybersecurity risks of weapons systems. It explores the three views of cybersecurity in cyberspace [55], [56] using a technical, social-technical and governance perspective. These views are combined with Dunn Cavelty's [57, Para. 1] theory of *threat representation.* Her theory brings together main actors, referent objects and threats of cybersecurity into three threat representations clusters. Results of the interviews are combined in this chapter.

The fourth chapter describes the current views on the possibilities of various forms of regulation. It highlights the different ways and gives a framework of possible strategies. In this part, the framework is analyzed using empirical evidence built up from interviews with stakeholders.

The final chapter will present the analysis and conclusions based on the research question presenting the most profitable strategy with arguments to succeed. It will focus on the underlying assumptions and conceptions of security concerning national security, international relations and the states' ongoing struggle to maintain or increase power.

2. The complex network of autonomous weapon systems

"The risks of functionally delegating complex tasks—and associated decisions—
to sensors and data-driven algorithms is one of the central issues of our time,
with serious implications across sectors and societies.
Nowhere are these more acute than in relation to decisions to kill, injure and destroy
- *Neil Davison* -[32], [58]

2.1. Weapon Systems

Some authors argue that the debate to define an autonomous weapon system is a semantic dispute [59, p. 937]. Legal scholars, computer scientists, and military, among others, have contributed to this debate each with their respective perspectives. First, it is possible to make a linguistic distinction between the terms 'weapon' and 'weapon system'. The former might refer to 'all arms, munitions, materiel, instruments, mechanisms, or devices that have an intended effect of injuring, damaging, destroying or disabling personnel or property' [60] while the latter is more broadly conceived to include 'the weapon itself and those components required for its operation, including new, advanced or emerging technologies [32], [61].

The Geneva Conventions, the Convention on Certain Conventional Weapons (CCW) and the Arms Trade Treaty (ATT), the Wassenaar Arrangement (W.A.) and the Missile Technology Control Regime (MTCR) all define weapon systems differently [32]. Defining an autonomous weapon in the legal context is important in order to proof legitimacy in terms of the law of armed conflict. It is obvious that an weapon system cannot be employed if it cannot apply with the law of armed conflict. Mainly the international debate is doubting whether autonomous systems could be capable of complying with the rules of international humanitarian law.

From a technical perspective a 'weapon system' is understood to be a system that may consist of multiple physical platforms, including carrier and launch platforms, sensors, fire control systems and communication links needed for a weapon to engage a target [62]. Although some weapon systems are purely I.T. systems, most—such as aircraft, missiles, and ships—are what the National Institute of Standards and Technology [36] and Zahid [62, p. 1] refer to as "cyber-physical systems." NIST defines these systems as "co-engineered interacting networks of physical and computational components [32]." When these systems might feature some autonomy in their critical functions, they can autonomously search for, detect, identify, select, track or attack targets we speak of an autonomous system [63]. The other components could also include other functions that might be autonomous, such as autonomous navigation, take-off and landing, or even flying. Modern weapons systems can contain hundreds of thousands of chips; each can be sophisticatedly designed, containing billions of transistors, that are functioning in a cyber-physical system, which generally involves sensing, computation and actuation. A cyber-physical system involves traditional information technology as in the transfer of data from sensors to the processing of those data in computation, it also involve traditional operational technology for control aspects and actuation. The traditional cyber-physical system was a closed serial network that contained only trusted devices with little or no connection to the outside world [45]. Older cyber-physical systems communicate data with operators over radio transmissions or are closed systems. But due to the innovative commercial communication possibilities and the upcoming technologies such as 5G real-time information exchange becomes more and more reality in the future [64], [65]. IoT technology might play an essential role in military action's effectiveness and contribute to its situational awareness capabilities. This interconnection is also defined as the Internet of Battlefield things (IOBT) [45]. Real-time battlefield data-sharing and the cooperative decision-making among commanders could become highly dependent on the connectivity between different combat units in the network [32]. An autonomous weapon system may perform some functions better from a military perspective (in some cases faster, stealthier, or more precisely) than human equivalents; this will also reduce risks to human soldiers. Evolving to this form an autonomous weapon system can become a highly complex systems-of-systems [27]. Automation in weapons technology is also inevitable as a response to the increasing tempo of military operations and political pressures to protect not just one’s own personnel but also civilian persons and property [23].

2.2. Autonomy in weapon systems

A broad range of weapons systems is in service [23]. Automatic systems have existed for decades. The first automatic air defence system, the Mark 56, was invented during World War II [27] and in the 1960s, the Patriot air [X2] defence system was designed. However, this system is still under human control: operators who can intervene at any time in case of an unwanted launch. Thus, automatic or autonomous functions in weapons systems are not new, but they have become increasingly more “autonomous”. Defining *autonomous* in the notion of autonomy in autonomous weapons is problematic because it seems to have a changing context [27]. The term autonomous appears to refer to all technologies (irrespective they are automated, autonomous, learning, self-learning or another type of Artificial Intelligence). According to Paul Scharre [27, p. 1], this is a misperception about autonomous weapon system; autonomy in the autonomous weapon system debate refers to the system's intelligence, but 'how intelligent the system is' and 'which tasks are performed autonomously' are different dimensions [66]. It is freedom, not the system's intelligence, that defines the autonomous weapon [66, p. 542]. Therefore, it might not be adequate to ask whether an autonomous weapon system 'works well' or 'is dangerous'. Instead, it is necessary to compare the characteristics of autonomy with the corresponding current human action it replaces or complements on a case-by-case basis.

To define autonomous weapon systems, we need to define autonomy as a relative notion: within and across relevant disciplines, be it engineering, robotics or computer science, most experts have a different understanding of when a system or a system's function may or may not be deemed autonomous. However, to understand the capabilities of the weapon, the concept of autonomy needs some clarification. From this perspective, *autonomy as technology* is seen as a function to achieve a task. The definition of Bills [67, p. 2] has a focus on the ability of a machine performing a task: “LAWS are robots used to deliver lethal force that possess near-human decision-making abilities" [68]. Thus an 'autonomous system' is a machine (hence robot) that, once activated, performs some task or function on its own. Automation in this definition seems to increase when more tasks are automated in a system. Unfortunately, this definition does not specify functions and suggests that automation has no levels.

Autonomy in systems could also be seen from a function perspective, and thus a more specific definition might refer more to the process within the system is used. The definition of Trumbull [66] defines the functions somewhat more: `A weapon that, without human intervention, selects and engages targets matching certain predefined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention [66].' This definition is also still not very precise. In this definition human- system collaboration is not mentioned, and it does not provide a possible level of automation from a technical point of view.

Descriptions of autonomy in weapon systems with the concept of a technological spectrum should move from remotely controlled systems on one side to autonomous weapon systems on the other. The definition from UNIDIR takes that argument into account: Autonomy in weapon systems is understood as a spectrum of capability 'moving from remotely controlled systems on one side to autonomous weapon systems on the other. Autonomy increases when moving along the spectrum from objects controlled by human operators from a distance (such as remotely piloted unmanned aerial vehicles) to automatic and automated systems to fully autonomous ones’ [21], [31], [69].

Heather Roff measures the level of autonomy in a weapon system based on three capabilities of an autonomous weapon system: self-mobility, self-direction and self-determination. Self-mobility capabilities allow a system to move by itself, self-direction relates to target identification, and self-determination indexes the abilities that a system may possess concerning goal setting, planning, and communication [21], [31], [69]. Still, the level of automation is not seen as a continuum. Scharre and

Horrowitz [21] define levels of automation more specifically. Scharre and Horowitz [21] argue that the level of automation can vary along three main axes or dimensions. These dimensions are independent. So, autonomy does not exist on merely one spectrum, but on three ranges simultaneously: first, the human-machine command-and-control relationship; second, the sophistication of the machine's decision making; and third, the type of decision-function that is automated. Scharre and Horowitz refer to this as the three dimensions of automation: (1) the relationship between person and machine, (2) the complexity of the machine, and (3) the type of function that is automated.

In the first dimension, the **Human-Machine relation-**autonomy refers to the relationship between the person and the machine. Machines need to stop after performing a function or task and wait for human a decision to continue. This concept is called the *human in the loop* [70]. The second form is *human-on-the-loop* [71, p. 3]. Within this idea, the system function might be problematic, or the human might only be monitoring and have possibilities to intervene when the system malfunctions, the human here has only supervision. Third, there is a *fully autonomous system*; in this concept, autonomy is not based on the system's intelligence but rather about performing the tasks given without the interference of human control.

Automation can refer to the **complexity of the machine**. This second factor is a utterly different way which might refer to *automatic* - almost mechanical responses - to environmental input. The term *automated* can also refer to the system's complexity and describes a more complex, rule-based system. Alternatively, speaking of complexity, the term *autonomous* might refer to machines that execute behaviour that entails some form of self-learning, emergent behaviour that is not directly predictable from the inspection of its code. Within this concept, the boundaries between degrees of complexity are vague, and it is hard where to draw the line between automatic, automated, autonomous and intelligence, but Scharre & Horowitz [72] argue that it is meaningless to call a system automatic or intelligent without specifying the tasks or function that is automated.

In the third factor, the type of **function of automating**, different decisions have different complexity levels and other risks. The question is: 'which function is operated by machine and which by the person?'. The engagement between human and machine that defines the degree of automation or human control is the *human out of the loop* [73]. Engagement-related tasks could include: acquiring, tracking, identifying and cueing potential targets, aiming weapons, selecting specific targets for engagement, prioritizing targets to be engaged, the timing of when to fire, manoeuvring and homing in on targets, and the detonation itself role [13], [74].

2.3. Future: Autonomy in weapons systems

In a RAND research report, Snyder, Powers, et al. [75] write that most modern U.S. military systems are so intertwined with cyberspace that they depend on it for their fundamental operations. Many are further connected, either directly or indirectly, to other military systems, forming a complex systems-of-systems whose capabilities are interdependent. State-of-the-art weapons systems are deployed beside legacy high-value systems. Many of the applications of such autonomous systems are in their infancy; the consequences of such accidents are probably limited. However, this is likely to change in time, not least if there is further convergence between the development of autonomous systems based upon machine learning and efforts to weaponize them [45]." Thus, it is unlikely that Artificial Intelligence technology such as Machine-learning will be fully applicable for complete weapon systems in the short term. Does this mean that advances in Machine-learning or deep learning are not needed to counter the ever-evolving threat posed by adversaries? In general, Artificial Intelligence is used to automate the detection of attacks and evolve and improve their capabilities over time [20, p. 3]. Buchanan et al. argue that a dual-use will be more realistic with a rule-based system or human input. Automation can also perform 'symbolic manipulations' that are very similar to what appears to be playing out in the human cognitive system, thus real-life thus the environment. Such machines, therefore, seem to show a certain degree of intelligence. The environment can be small or large, specific or very wide. The definitions "narrow" or "general" Artificial Intelligence systems have their foundation on this principle. The most crucial distinction between alternative Artificial Intelligence systems is thus generality. Today Artificial Intelligence systems can only solve narrow sets of tasks in limited environments, even though Artificial Intelligence already defeat humans while playing chess or poker, in language processing skills or face recognition possibilities and surveillance. In other different areas, the human is still superior; in general, the conclusion is that the more depending on the task's complexity, the more human control or command is needed. Until today, the systems remain narrow, and it still cannot generalize as broadly as human can [4, p. 18]. Thus, Artificial General Intelligence is more of an idea in the future than in today's reality [40]. Thus Artificial Intelligence is not yet entirely suitable for use in complex environments.

2.4. Sub-conclusion

The increased aspirations of the application of artificial intelligence, and the conceptual foresight that an autonomous weapon system might perform functions better from a military perspective - faster, stealthier, or more precisely- than human equivalents brands an autonomous weapon system as a highly complex systems-of-systems [41]. However, the use of autonomous weapons is generally considered revolutionary and legally under debate. The discussion about autonomy as a general attribute is imprecise. One of the major challenges is the conflicting descriptions of autonomy from either a general sense, or in reference to a particular context or system. Autonomy in weapon systems can serve different functions – like targeting, flying, navigation - and the degree of autonomy might depend on the use of specific tasks. Therefore a definition will benefit from a more platform or system-centric approach. Within this system-centric approach, the element of human-machine and thus the possibility of meaning full human control defines the autonomy in the system especially in the targeting function of the weapon itself. The level of autonomy is not inherent to a weapon system, but determined by what it is allowed to do without human oversight or control. Form a technical point of view the targeting function relates to the intended effect of injuring, damaging, destroying during conflict, independent from the defensive or offensive military objective of the mission within the weapon is functioning. The definition of autonomous weapon systems can be more specific if it entails also the political, legal and military decision making process that takes place on before, during and after the actual occupation of the weapon, the risk of an unintended casualty starts here. Nevertheless these, by human predefined criteria, are deployed in a cyber-physical systems that needs to be secured to perform as intended.

3. Cybersecurity risk factors of weapon systems.

"War is dangerous,
and weapons that are intended to be deadly to an opponent
often can be quite dangerous to the user or friendly forces as well [42], [43]."

As the authors of the 2018 published report of the American Government Accountability Office wrote: "Due to the lack of focus on weapon systems cybersecurity, the Department of Defense likely has an entire generation of systems that were designed and built without adequately considering cybersecurity [20]". Jahan et al. [33] also conclude that autonomous systems' cybersecurity is at a very early stage. Weapon systems are just like information technology (I.T.) or operational technology (O.T.), vulnerable to cyberattacks. Those operational risks, such as those related to the reliability [34], fragility [76], and security of systems [34] could raise fundamental questions about whether military systems will function and operate according to military commanders' intent [35], [36]. The threat that malicious users hack or control such systems by exploiting their vulnerabilities could be realistic [77]. The 2009 theft of the F-35 fighter designs from the US military through Chinese hackers is one of the most high-profile cases of cyber espionage [56]. Even though often used network security measures, such as air-gapping, ensure that a weapon system lacks a direct connection to other untrusted systems networks, attackers might exploit external interfaces (e.g., radios, radars, and maintenance ports) [6]. Hackers might try to gain access or manipulate weapon systems' internal computers, networks, and data [70] which may result in security implications [70, p. vii]. Implications affect the system itself and the tactical and operational value that a military commander expects when the system is used. Interviewee [X1] referred here to an important military principle: "In the military, it is safety always, mission first. Cybersecurity is one of the different tools we use." Cybersecurity issues seem to affect the system itself and emerge as a political consideration for states, multilateral organizations, firms and civil society [78]. However, the term cybersecurity is widely used and has evolved rapidly while convergence with other forms of security [79]. The desired outcomes of military cybersecurity management are, following the description of Snyder, Powers, et al. [80]: "to limit adversary intelligence exploitation through cyberspace to an acceptable level and to maintain a proper operational functionality (survivability) even when attacked offensively through cyberspace" [81].

3.1. Security of Cyberspace

The exact meaning of cyberspace is inadequately and unclearly defined in the literature. There are different views on the context, domains of application and terminologies [82]. From a military operational perspective, cyberspace refers to the fifth military domain, next to sea, land, air and space [80], but that classification does not describe the function of cyberspace. Cyberspace can also be seen as a place where governmental, military, commercial and private users make connections via the internet and social media [80]. Cyberspace is unique as it also contains virtual, more or less ethereal, elements [83], and cyberspace is unlike the air, space, or the sea, an entirely man-made area. The US Army understands cyberspace intended and designed as an information environment [83, p. 24], although they acknowledge cyberspace's expanded appreciation today. The US army [84] describes cyberspace as a three-layer model with five sub-layers consisting of a physical, logical, and a social layer comprising the following five components: 'geographic, physical network, logical network, cyber persona and persona' [85]. Valeriano and Maness [86, p. 1] define cyberspace "as the networked system of microprocessors, mainframes, and basic computers that act in digital space. Cyberspace has physical elements because these microprocessors, mainframes, and computers are systems with a physical location. Therefore, cyberspace is a physical, socio-technological environment that interacts and blends with other domains and layers [86, p. 1]." This definition focuses on what the term cyberspace entails and indicates the problems that might result from it, but it places no normative judgment on the efficacy of cyber issues, something Van den Berg's model does. Van den Berg's [55] model considers that indeed cyberspace involves a technical and socio-technical layer, but that it also involves a third layer: the Governance-layer. This layer exists above the two other layers and should govern cyberspace through regulation, rules or standard-setting norms [87]. Cyberspace in that notion can be seen as a dynamic, evolving, multilevel ecosystem of physical infrastructure, software, regulations, ideas, innovations, and interactions influenced by an expanding population of contributors [5].

Cyberspace can be seen as the operational environment where threats and risks could occur and where thus, security measures could help protect cyberspace. A further operationalisation of cybersecurity is needed to understand what the notion of cybersecurity involves. Lewis defines cybersecurity as "the safeguarding of computer networks and the information they contain from penetration and malicious damage or disruption." [77], this definition is unclear about what safeguarding should entail, and thus the outcome, the actors, and the structure are vague. According to Van den Berg [88], [89], when cyberspace was originated, the security focus was on information security with the basic concepts of confidentiality, integrity and availability of information, the CIA-Triade. From this point of view, cybersecurity can be described to be the set of technologies, processes, and practices designed to protect networks, computers, programs, and data from attack, damage, or unauthorized access, by the common information security goals: the protection of confidentiality, integrity, and availability of information [88, p. 11]. Nevertheless, the classical definition of information security is perceived as too restrictive for cybersecurity [85]. Van den Berg [90, p. 7] argues that non-technical issues like economic, ethical, legal and governmental aspects are also part of cybersecurity. The traditional CIA-Triade ignores the semantics between the interactions of actors and the IT systems. Cybersecurity issues seem to affect systems' defence and demand political consideration or intervention for states, multilateral organizations, firms and civil society [90, p. 7]. Thus “cybersecurity requires not only protecting and defending society and its essential information infrastructures but also a way of prosecuting national and international policies through information-technological means [91].” Security in this point of view is not something you have; it is as a threat to (core) values of society [90, p. 7].

3.2. Defining risk factors

Deibert & Rohozinski [85] divide risks for securing cyberspace into two related dimensions: risks to the physical realm of computer and communication technologies (risks to cyberspace); and risks that arise from cyberspace and are facilitated or generated by its technologies but do not directly target the infrastructures per se (risks through cyberspace). Hansson [89] observes that risk can mean an unwanted event, which may or may not occur [89] and that ‘risk’ is assumed as the cause of a potential unwanted event. Furthermore the complexity of the risk is frequently used interchangeably with notions such as hazard, threat, loss, and damage [89]. In military operations this notions are critical to identify, because how critical a system vulnerability might is depends on its use to support operational missions and which missions it supports [89], [92]. The appropriate level of risk of critical mission systems are established by the so-called mission assurance risk approval. Mission assurance is, according to MITRE[1]: “the ability of operators to achieve their mission, continue critical processes, and protect people and assets in the face of internal and external attack (both physical and cyber), unforeseen environmental or operational changes, and system malfunctions.” Securing cyberspace or an autonomous weapon is a complex task, and risk assessment can reduce that complexity [X4]. It reduces the complexity of a particular situation by assessing the impact of – often undesirable – events into a chance or probability that a certain incident will occur. The purpose of risk calculation is that the statistical expectation value of unwanted events may or may not take place. Risk calculation involves thinking of risk as the probability that an unwanted event may or may not take place, and it also entails identifying and calculating the consequences of that event. This process will lead to the identification of potential incidents that should be mitigated or accepted when the mission risk approval certification is performed.

[1] https://www.mitre.org/publications/systems-engineering-guide/enterprise-engineering/systems-engineering-for-mission-assurance#:~:text=Share,operational%20changes%2C%20and%20system%20malfunctions.

An essential element in deciding how to handle risks is finding ways to cope with an uncertain future and prevent or minimize possible losses [93, pp. 54–55]. The assumption that underlies this idea is that we can influence future outcomes and keep ourselves safe by taking appropriate measures. There are mainly three perspectives on how to cope with that uncertainty of taking the appropriate measure. Those perspectives are referred to as the realist perspective, the social constructivist perspective and the psychometric perspective[89]. First, in the realist perspective, risks are real events or hazards in the world and are determined objectively [94]–[96]. The realist perspective views risks from a technical, natural science, and engineering outlook. In this context, risks are considered real objective dangers or hazards identified through risk management practices. Risk management entails the process of objectively calculating and quantifying the likelihood and impact of events [90, p. 7].

The social constructivist perspective holds that risks cannot merely be determined by an objective calculation or measurement of an event's probability and impact. Instead, the risk is seen as a complex construct created by cultural or social values applied to it [56]. Thus, what constitutes risk is influenced by human interaction, involving social and cultural processes that steer the appreciation and determination of risk [56]. As opposed to the realist view on risk, calculating probabilities in terms of security is much more difficult. Human behaviour is not static, and malicious human actors' capabilities are often unknown [56]. First, the complexity of risks can be explained by looking at cyberspace's interconnected nature with the physical domain. Second perceiving an unwanted event as a risk is a decision that is made under conditions of known probabilities [56]. Consequently, the term risk denotes the possibility that an undesirable event may occur as a result, in the case of a cyber-activity, after a human decided to start that activity that may lead to that undesirable event.

The start of an unwanted event may be driven from a political point of view. Building on the notion that risks are political is a psychometric approach. The psychometric perspective fits in between the realist and social constructivist perspective since it studies the perception of risk. Since risks are not only a matter of choice -within certain boundaries-, risks also come with political and social effects [91]. Zooming in on the strategic-military aspects of cybersecurity might mean that it is subjected to the rules of an antagonistic zero-sum game, in which one party’s gain is another party’s loss [56]; this might lead to risk reduction too firmly focused on national security measures instead of economic and business solutions since it is wrong to suggest that states can establish control over cyberspace [56]. In all, this creates an unnecessary atmosphere of insecurity and tension in the international system, which is based on misperceptions of the nature and level of cyber risk and the feasibility of different protection measures in a world characterised by complex, interdependent risk [56].

3.3. Technical, cyber-crime and espionage, and cyberwarfare threats and risks

Risks are also defined by assessing *threat representations* [97] – different ways to depict what counts as a threat - to a weapon system. Dunn Cavelty [92], [98] argues that there are three discourses of cybersecurity threats relevant for the military domain: *technical cybersecurity, cyber-crime and espionage, and cyber-war*. In the first category, threats arise from malware and system intrusions or disruptions [99]. Within the second cluster, threats arise from cybercrime and cyber espionage actions. It makes a difference, from a legal and political point of view if a computer intrusion (technical threat) is attributed as a criminal act or as an act of espionage, resulting in using other response methods such as misleading, disturbing retaliation or prosecuting the identified actor. The third representation is cyberwarfare. The main actors here are nation-states' cyber force focusing on military networks or their systems [100]. In the following section the main threats and risks will be discussed categorized by the former described *threat representations*.

Technical cybersecurity

Within the technical cybersecurity representation, the **information security threats** are an important part. Researchers have proposed various theories and approaches to manage the threats by analyzing its risks [30], modelling its security [29], developing security strategies and policies [23], and establishing international security standards [6]. ISO [62], NIST [101]. Yeh [4] defines four main threats: interruption, interception, modification, and fabrication. These threats lead to the risk of not protecting information systems against unauthorized access to or modification of information, whether in storage, processing or transit, and against the denial of service to authorized users, including those measures necessary to detect, document, and counter such threats. Other research based on information security makes use of classification methods based on threat impact such as: spoofing identity, tampering with data, repudiation, information disclosure, and elevation of privilege [102], The NIST framework for cyber-physical systems defines security aspects and concerns which should be taken into account in the design process [103]. Carter et al argue from a social constructivist viewpoint, that is difficult to address security early in the design process and then continue during the lifecycle as both threats and the systems evolves [103]. They argue that security solutions are developed on perceived threats to system, this means that vulnerabilities are discovered after a breach or failure. Risk analyses could also be based on structured argumentation using attack-three methods like STRIDE: Spoofing, tampering, repudiation, information disclosure, denial of service, elevation of privilege to discover threats[103]. Designers are using these methods due to the unpredictability of attackers and zero-day vulnerabilities. Carter et al. also argue that traditional security requirements techniques are insufficient for cyber-physical systems due to the physical and component interactions inherent to such systems[104]. Young and Leveson showed that the cyber-security problem needs to change focus from responding to threats towards controlling vulnerabilities[103]. This refocusing to a top-down approach enables analysis to maintain a system-level perspective while also having the ability to focus in on loss scenarios in high levels of detail [105].

The threat of interruption is also important to discuss more in depth. **(Unintended) System Failures or mistakes** A considerable risk to autonomous weapon system is a *failure. A* failure refers to actual or perceived degradation or loss of intended functionality or inability of the system to perform as intended or designed [62]. These failures do not necessarily mean that the system fails to function when the risk of failure is assessed during the engineering of the system. An autonomous weapon must always meet the critical assurance requirements since malfunctioning could jeopardize the operation, the unwanted effect is that the system loses their mission readiness. A system therefore needs resilience to handle a failure when it does unexpected appears. Failures will therefore always stay a threat. Which may arise from cyberattacks, infiltration, to jamming, spoofing, software coding errors, human error, human-machine interaction failures, malfunctions, communications degradation, infiltrating in the industrial supply chain and unanticipated situations on the battlefield.

The NIST framework for cyber-physical systems also mentions concerns around **data** [101]. Data bias is a problem that can arise when data is modified. Threats to data are the loss of data integrity, data confidentiality, nonrepudiation, and data freshness. Occurring biases are (self)selection bias, exclusion bias or reporting bias and detection biases[102]. Misrepresentation can lead to a vicious circle and may lead to a disruption of the system. This is could be assessed as risk that can lead to modification and maybe to a system failure, but the formal validation of predictability stays a challenge [101]. This risk can also be viewed from a social constructivist perspective. It is the human influence on the data quality and the human primary input to build data sets on which the systems are build. The mitigating approaches should tackle bias in different stages of autonomous decision making, namely, pre-processing, in-processing, and post-processing, but human testing of the results is essential. However, autonomous decision-making may enlarge pre-existing biases and evolve new classifications and criteria with massive potential for new types of biases [28]. Data security is also mentioned in the interviews [X2, X3, X6, X7] as a risk that should be identified. Due to the increasing connectivity level, a permanent focus is needed for confidential, and thus encrypted communication of data. The transport of the data is seen as threat. Simultaneously, the interviewees are apprehensive about the required quality and amount of data, especially with the future of increasing autonomy in mind. With a higher level of autonomous actions of a system, the need for the system's intelligence will also rise. Intelligence has to rely on data-sets; however, therefore, there must be enough data with the required quality. This concern is also debated in the literature when there is referred to the BlackBox. Machine learning techniques might help to enable or improve the performance of systems like an unmanned combat aerial vehicle, a marine vessel surveillance, or for detecting and mapping mines. Those systems need large amounts of sensor and intelligence data to support military decision making. Challenges are that this systems are very complex. Autonomous weapon systems might require novel approaches towards explainability, since the operation of many of these machine learning models is often challenging to interpret by the designers, other actors and users: the greater the volume, variety, and velocity of data processing, the more difficult it becomes to understand and predict a system's behaviour or re-construct the computed correlations [73]. This re-construction problem is severe, because even with complete information about a system's operations, an ex-post analysis of a specific decision may not be able to establish a linear causal connection that is easily comprehensible for human minds [106].

The unwanted event is that operators, designers, and the user cannot interpret, trace and check the trustworthiness of the Machine Learning or other AI algorithms. The systems need to communicate clearly and openly with some level of trust. Accomplishing this trust-building through understandable and clear communication is a challenge, and therefore, the explanation mechanisms involved should allow for two-way information flow between different parties involved in the design, implementation and exploitation phases [107]. In general, humans are careful to adopt techniques that are not directly transparent. However, there is a trade-off between a model's performance and its transparency. The risk is that the trade-off performance wins over transparency to favour the end-user [108].

Cyber-crime and espionage

The second threat representation cluster focuses more on the threat of unauthorized control of an autonomous system, theft, data poisoning and disturbing. **Cyber-attacks** can harm a system and thus the risks losing control, system failures via modification or exploitation of an autonomous weapon system arise [15], [36], [60], [61], [63]–[65]. The terms intrusion, penetration, attack, breach and compromise are often used interchangeably when scholars discuss cyber-attacks [37], [109]. An intrusion (or penetration) can be seen as a successful event from the hacker’s perspective and an unwanted event in a defender's point of view. An attack in a vulnerability can be exploited, and a breach in a system results from a violation of a (security) policy. An attack that does not lead to a breach can be considered unsuccessful, although the attacker might have access to information. Lindqvist et al. [4] classification gives three main categories of intrusion techniques: *bypassing intended controls*, *active misuse of resources* and *passive misuse of resources*, which may result in the following unwanted events: *exposure of information, denial of service, incorrect output*. The basic idea with incorrect output is to fool a system by providing malicious input, which often involves minimal perturbations that cause the system to make a false prediction, categorization or decisions [105]. This can lead to a another important risk **undetected attacks.** Not every incident is detected; a fraction always goes undetected [110]. Therefore, detecting measures should not only focus on the unusual behaviour of a system but also the deviation of normal behaviour. For cyber exploitation to be successful, an adversary needs to gain access to useful information and exfiltrate it before it is detected and blocked [76]. Access can be gained through software infiltration or by implants in firmware or hardware introduced through the supply chain. These access points can, in principle, be indirect via a less critical system that has some information exchange with the system containing the more critical information [76], [101], [102]. Hence, to have effective counter cyber exploitation, the system owner needs to identify the most critical information and the adversary needs to be denied access to that information; if the adversary gains access to critical information, he must be detected and blocked from successfully exfiltrating it [100].

Interception of connection. Interviewee [X1] argues equally as the American Government Accountability Office [111] that any information exchange is a potential access point for an adversary. Even "air-gapped" systems that do not directly connect to the Internet for security reasons could potentially be accessed by other means, such as USB devices, connected to a system. Weapon systems have a wide variety of interfaces, which all potentially could be used as pathways for adversaries to access. Historically experts were focused on the cybersecurity of networks but not weapon systems themselves [56]. Interviewees [X1] and [X6] have indicated that this former perception is not fully recognized, most weapon systems have measures against untrusted connections, and their architecture is highly compartmented. Nevertheless, both interviewee's endorse the claim that the risk exists.

Cyberwarfare

The third perspective under discussion is the threat representation of cyberwarfare. By assessing the risk a difference must be made between (domestic) law-enforcement operations, (national) intelligence operations and cyberwarfare operations, while the performed technical cyberactivity may be equal, the ends and means may be different and thus the perception of risk may change. Cybercrime and espionage apply to other laws than cyberwarfare. Cyberwarfare have sparked extensive debates in regards to compliance to current international law, and scholars have arguments on both sides to what extend cyberwar could take place, but that point will be discussed comprehensively in the next chapter. Cyberwarfare, as an unwanted event, may lead to risks for the functioning of the cyber-physical system itself, for the operation, for the operators or even for society as well. Therefore this type of risks not only has an impact on the technical layer, but also on the socio-technical and governance layer. In the governance layer of cybersecurity, attribution can be a mitigation against the risk of Cyberwar. Attribution on the root cause of attack might or might not reveal the actor, this relates to the inherent difficulty of attributing cyber conduct, which may of course be directed via identity masking (or spoofing) tools and similar electronic misdirection, to a specific perpetrator or actor. This can be explained through the lens of the psychometric perspective [53]. In cybersecurity terms, through the model of van den Berg, an autonomous weapon is a socio-technical activity within cyberspace that, if attacked by an adversary, can cause unintended physical harm, since the attack on the system might lead to psychical harm -or even lethal consequences- to society, military personal or it has a negative effect on the performance of the military operation. But in there a certain caveats to cyberwarfare: the consequences of the technological possibilities to attack are mediated by the ability of state driven military cyber commands to use them effectively in relevant military scenarios and in the pursuit of political ends.

3.4. Sub-conclusion

Cyberspace can be seen as a dynamic, evolving, multilevel ecosystem of physical infrastructure, software, regulations, ideas, innovations, and interactions influenced by an expanding population of users. The vulnerabilities, threats and risks that are identified in this chapter showed that they can apply to cyberspace itself when involving the physical realm of technologies, or showed that these issues can arise from cyberspace and are facilitated or generated by the use of technology [90].

One of the cybersecurity challenges in achieving secure autonomous weapon systems is the perception of the risks based on real objective dangers and hazards. These complexity arise in the phase of the technical design, and evolves during the introduction of emerging technology and changes continuous due to the interconnected nature with the physical domain. The complexity of the systems combined with the need to deliver a weapon system that will function as wished for, makes functionality of the weapon system intertwined with cybersecurity. The detailed engineering operations of the protocols, the never ending process of identifying critical vulnerabilities, and understanding how to address these vulnerabilities without compromising functionality, are so complex that specialists continually need to make trade-off in order to achieve functionality by accepting a certain level of vulnerabilities. This risk acceptance ought be in accordance with international law, regulations which are present in the mission objectives of every legal military operation. By the best effort, the research performed for this thesis did not found a risk assessment framework that involves this acceptance process.

Second autonomous weapon systems need to apply to an appropriate level of risk, depending on the task that it needs to deploy in the military operation where it legally participates. The identification of basic information security risks showed that standards and frameworks for securing exists. The ethical discussion around the possible lack of transparency, the BlackBox, is key point of debate, combined with the risks of data bias more research is needed to investigate to what extent this can lead to a non-acceptable risk, but may also be hard to mitigate completely. The need to have full control over data makes it important to develop requirements for making algorithms secure, reliable and robust with a trusted community within the ecosystem. This investigation did not examine in more detail the extent to which this transparency gives the owner of the system new possibilities for control.

4. Regulating Autonomous Weapon Systems in Cyberspace.

Cyberspace can be seen as a dynamic, evolving, multilevel ecosystem of physical infrastructure, software, regulations, ideas, innovations, and interactions influenced by an expanding population of users. The vulnerabilities, threats and risks that are identified in the former chapter showed that they can apply to cyberspace itself when involving the physical realm of technologies, or showed that these issues can arise from cyberspace and are facilitated or generated by the use of technology [90]. Risks mitigating is a central element in the field of cybersecurity. This chapter will investigate the possibility to what extent it is possible to define strategies of regulation in order to mitigate the risks defined in chapter three.

4.1. Strategies of regulation

In general, regulation involves three key aspects: standard-setting, behaviour modification and information gathering [28]. The three aspects are separate, but their function is interdependent and seen as the regulatory regime. *Standard-setting* is seen as 'the rule' itself and sets out the objectives or choices to the military organisations and military industry whose activities are regulated [112]. State-based agencies or self-regulatory bodies can do this, and within this aspect, regulation is seen as a 'toolbox' with various techniques and alternatives to address policy problems for building an effective strategy. *Behaviour modification* is the overall intended behavioural change why rules are composed and enforced. Enforcement makes 'rules happen', so achieving compliance is possible through advice or threat and persuasion. *Information gathering*; elements of detection are necessary. Enforcement or auditing on standards is not potential when there is no information about monitor activity that provides information.

Regulation can be precautionary, preventive, responsive, or a combination of all three. Responsive regulation rests on the core principle that escalating enforcement practices are deployed when the regulated actors’ behaviour does not comply [113]. In this context, the default strategy is non-intrusive and delegated regulation, which is more likely to generate cooperation and innovation among private actors by allowing them discretion in deciding how best to achieve regulatory goals.

A second option is to pre-emptively limit or even ban certain applications out of fears for worst-case scenarios. This is known as the “precautionary principle.” The precautionary principle generally refers to the belief that new innovations should be shortened or excluded until manufacturers can prove that they will not cause any harms to civilians or various existing laws, norms, or traditions. The concerns of new risks have led to a call for precautionary regulation. But prior restraints on innovative activities are seen as recipe for economic and social stagnation [114]. Scholars argued that precautionary regulation was the process of “creative destruction [114]” others claimed that the risks reduced and business became on an innovative edge and competitiveness via regulation. The fundamental weakness in these claims is that regulation here is seen as one kind of type or rule, with only one effect[115]. In practice, regulation is like the technology itself; it comes in various types and has various effects. Different regulatory designs can delay or accelerate technological change, or shape it in varying ways, favouring some kinds of technology over others [28].

The third option is to develops preventive measures and policies to protect on forehand, standard-setting is a preventive measure that is commonly used. Providing national security and the control of legitimate force is the essence of the modern nation-state and differentiates it from functioning as a failed state [102]. Thus, the nation's strategy includes ideas on making the state efficient using instruments of power.

4.2. Strategy for technical risks and data

The identified risks for autonomous weapon systems based on the NIST framework for cyber-physical systems, which defines security aspects and concerns that ought to be taken into account in the design process [100]. Well respected international standards bodies, ISO and IEEE, have an extensive history of governing a range of socio-technical issues. Standards bodies and their processes facilitate the arrival of consensus on what should and should not be within a standard. Engineers and organizations have the possibility to use standard from established international security standards come from ISO [62], NIST [4] or IEEE, but nevertheless, the American Government Accountability Office [4, p. 18] wrote that there is a lack of focus on weapon systems cybersecurity, stating that “bolting on cybersecurity late in the development cycle or after the system's deployment is more difficult and costly than taking security into account during the design[103].”

Carter et al. point out two arguments why working with security standards is complex. First they that argue that is difficult to address security early in the design process and then continue during the lifecycle as both threats and the systems evolves [103]. They argues that security solutions are developed on perceived threats to system, this means that vulnerabilities are discovered after a breach or failure. Second, Carter et al. argue that traditional security requirements techniques are insufficient for cyber-physical systems due to the physical and component interactions inherent to such systems[104]. Young and Leveson showed in their research that the cyber-security problem needs to change focus from responding to threats towards controlling vulnerabilities [103]. This refocusing to a top-down approach enables analysis to maintain a system-level perspective while also having the ability to focus in on loss scenarios in high levels of detail [114].

There are more external factors which may help to identify why working with security frameworks or standards are complex. Blind [114] describes that standardization is a voluntary process for developing technical specifications based on consensus amongst the interested parties themselves: industry in the first place, interest groups and public authorities. Cyberspace as industry is a private marketplace and the military industry are largely and highly private marketplaces. Military industry sells most of their production to their own government or friendly governments of their own State, using export control. Since the public authorities are also involved, this means that the governments and markets has several different incentives and interests. According to the network externalities-theory [28], parties need an positive incentive to cooperate, maintaining the standard in practice cooperation then becomes essential, as the harmony of interests obviates enforcement. Others roles might be more political and economic such as stimulating innovation or creating security through exports control of technology.

The call for transparency and thus, the mitigation of the BlackBox might be achieved via regulation. This regulation should aim at two things, according to Wischmeyer & Rademacher [28]: First, to provide different groups of stakeholders—citizens, consumers, the media, parliaments, regulatory agencies, and courts—with the knowledge each of them needs to initiate or conduct an administrative or judicial review of a decision or to hold the decision-maker otherwise accountable. Second, transparency can and should impart a sense of agency to those directly affected by a decision to generate trust in the decision-making process. By requiring system operators to explain their decisions to those affected, transparency is enforceable. However, to explain the BlackBox some regulation concerns come along [106]. Erdélyi and Goldsmith [106] call for a global regulatory body. They see a growing legal vacuum, created through issues like the externalities of transcending national boundaries where domestic approaches tend to conflict with the transnational nature of A.I. This discrepancy raises significant difficulties between the national law governing and it creates pressures for international regulation. By designing international law from the beginning, as argued by Erdelyi and Goldsmith [116], the risk of fragmentation is mitigated. It prevents creating too many rules of national interest at the beginning. Regulation, in this sense, could help the government achieve its own goals in providing national security. For the legitimacy of the state, citizens must trust the supervision and appropriate use of technology by the government and military.

4.3. Cybercrime, cyberespionage or cyberwar

Despite the fact that an attack through cyberspace will might lead to the same results a difference must be made between (domestic) law-enforcement operations, (national) intelligence operations and cyberwarfare operations. The reason is that cybercrime and espionage apply to other laws than cyberwarfare. Defining how cyberwarfare how it articulates with current law is still heavenly under debate, opposite of cybercrime, where the Budapest Convention (2001) aims to harmonize national cybercrime legislation, enhance transnational policing measures in pursuing and prosecuting cybercriminals, and improve international cyber-crime cooperation [116]. The debate about cyberweapons and cyberwarfare rests on the interpretation of existing international law [53].

To determine whether a cyber activity amounts to use of force and can be classified as an act of cyberwarfare, the -legally non-binding- Tallinn Manual can assist [112]. It has identified eight factors and among those factors are the levels of severity, invasiveness, and State involvement [112]. As regards severity, the manual specifies that cyber activities resulting in physical harm to persons or property may qualify as an act with the use of force [112].

The reluctance of governments to agree on and enact legally binding rules at the global level, less formal, norms-based discussions have emerged as alternative pathways to formal regulation. In contrast to binding legal statutes, norms as The Tallinn Manual refer to a voluntary standard forms of behaviour.

5. Increasing Cybersecurity of Weapon systems

5.1. Analysis

This thesis focuses on the extent to which a regulatory intervention can help secure an autonomous weapon system from the latest cybersecurity threat. The discussion about autonomy as a general attribute is imprecise. One of the major challenges is the conflicting descriptions of autonomy from either a general sense, or in reference to a particular context or system. Autonomy in weapon systems can serve different functions – like targeting, flying, navigation - and the degree of autonomy might depend on the use of specific tasks. Therefore a definition will benefit from a more platform or system-centric approach. Within this system-centric approach, the element of human-machine and thus the possibility of meaning full human control defines the autonomy in the system especially in the targeting function of the weapon itself. The level of autonomy is not inherent to a weapon system, but determined by what it is allowed to do without human oversight or control.

Automated (or semi-autonomous) systems may operate without a human, yet remain within the sphere of human control as they act according to pre-programmed specifications. However, autonomous functions should always be understood in relation to human action, and it is decisive in what context these functions are used. Autonomy means not that the operator's loses control on the tasks or functions that the system has; it means that the human tasks takes a different form. Human control can take various forms that can be considered meaningful or not in conjunction with the context. The human's accountability, on legal grounds, are not different. In the debate it is often wrongly interpreted that autonomous weapon systems determine their own goals, and it is not correct to frame autonomy in the sense of 'making decisions without being controlled a by human'. Yet these systems are in fact bounded by the tasks and/ or goals assigned within the legal framework of the mission of which the weapon is part. It is not realistic to suggest that human control disappears. Secondly and equally important, it may not be ethical or legal as well. At the same time, it is in the interest of humanity to recognize the use of war as an extraordinary remedy, from that point of view, a debate on using weapons in a generic sense through regulation helps to set the limits on the use of weapons by creating norms. Norms regulate behaviour in cyberspace, and norms may help achieve more secure cyberspace.

Adversaries are also developing new cyberweapons from which new threats arise, these new forms of uncertainty call for a continuous recalibration of resources to protect the ever-changing threat. As cyber threats become ubiquitous, the need to maximize the security level of systems rises. Interviewees watch attentively to speed on which other nations states are innovation and how they may use new technology in future conflicts. Therefore the interviewees argue it is vital to innovate, and dually create norms about autonomous warfare and the future conflict. Innovation is necessary to anticipate and understand the opponent's moves to have an adequate response. The innovation of autonomous systems and artificial intelligence is inevitable, because it is not merely a feature of weapons technology, but also part of technology innovation in general. Automation in weapons technology is also unavoidable as a response to the increasing tempo of military operations and political pressures to protect not just one's own personnel but also civilian persons and property [114].

The literature shows that automation seems inevitable in many areas, automation in weapons most likely will occur incrementally. All Interviewees see emerging technologies such as artificial intelligence in future military capabilities. However, they are also realistic that the path of innovation, procurements and training and exercise will take decades to. Interviewee [X4] indicates firmly that at this moment, relatively simple actions from people are taken over by systems, where mainly rule-based narrow artificial intelligence is used, the interviewee argues that the trade-off to entirely rely on Machine-learning will not happen and may be overhyped. The interviews with experts [X4, X6, X7] portray an comparable reaction. Interviewee [X4] argued that much of the developments are actually about automation not so much about autonomy. Autonomy in weapons systems can positively promote a war strategy's objectives through technological configurations and improving operational conditions, but not in all cases. The military operational viewpoint of scarcity can already explain this increase during operational conditions; there will always be a need to increase military capabilities and accelerate the speed of action favouring the adversary.

The discussion about the safe-state of the systems during the life cycle is much less conducted. Elderly weapons systems' service lives sometimes are extended or their performance enhanced due to reduced budget or setbacks in new procurement. Norms could help address this life cycle problem by setting a commonly accepted standard between a military organization and its suppliers. It ensures that the control of the autonomous weapon systems remains at the responsibility of the military organization.

One of the cybersecurity risks in achieving secure autonomous weapon systems is the perception of the risks based on real objective dangers and hazards. These complexity arise in the phase of the technical design, and evolves during the introduction of emerging technology and changes continuous due to the interconnected nature with the physical domain. The complexity of the systems combined with the need to deliver a weapon system that will function as required, makes the functionality of the weapon system intertwined with cybersecurity. The detailed engineering operations of the protocols, the never ending process of identifying critical vulnerabilities, and understanding how to address these vulnerabilities without compromising functionality, are so complex that specialists continually need to make trade-offs in order to achieve functionality by accepting a certain level of vulnerabilities. This risk acceptance ought be in accordance with international law, regulations which are present in the mission objectives of every legal military operation.

Autonomous weapon systems need to apply to an appropriate level of risk, depending on the task that it needs to deploy in the military operation where it legally participates. The identification of basic information security risks showed that standards and frameworks for securing exists. This form of self-regulation, with market norms, frameworks, principles is in place and seems to function.

The ethical discussion around the possible lack of transparency, the BlackBox, is key point of debate, combined with the risks of data bias. The need to have full control over data makes it important to develop requirements for making algorithms secure, reliable and robust with a trusted community within the ecosystem. This investigation did not examine in more detail the extent to which this transparency gives the owner of the system new possibilities for control.

The domain for autonomous weapon systems is an innovative niche market, making it a complex environment and a military organization is heavily depended on that market. It is this complexity that is the real enemy of cybersecurity. According to the network externalities-theory [54], parties need an positive incentive to cooperate, maintaining the standard in practice cooperation then becomes essential, as the harmony of interests obviates enforcement. Developers and engineers in this market should be encouraged to collaborate and share knowledge, incentives to encourage this knowledge exchange with a trusted military environment can contribute to the development of necessary knowledge of these innovations within the military, since gaining more access to knowledge is in the interest of the military command to stay in control over this innovation. The complexity and scale of innovation make it impossible for the smaller military organizations to innovate independently, but close cooperation, via standard-setting and norms the market can be steered towards a secure development.

Cyberattacks are becoming more frequent and it is therefore certainly not excluded that opponents will try to gain access to the locations where autonomous weapon system is being developed through espionage or want to gain access to the weapons system itself. From a cybersecurity perspective, a cyberattack is always an undesirable activity, the cause of which is either criminal or condoned in the case of espionage, but for protection from the perspective of cybersecurity is never tolerated. Traditional security measures are needed to create an initial blockade. But there are also other means such as trying to change the behavior of nations by developing together norms and laws to slow down the arms race. Governments must work together and agree to enact legally binding rules at the global level, since the less formal, norms-based discussions have emerged as alternative pathways to formal regulation.

5.2. Conclusion

Autonomous weapons systems will significantly increase the strategic advantage for militaries and should be designed to function under meaningful human control. As argued in this thesis, regulation mitigates the risks that nations will have an incentive to rapidly acquire and integrate systems without placing appropriate cybersecurity policies to ensure that systems are safe and reliable [54]. Regulation is a tool to support governance [54] and can apply to various forms of inquiry shaping social behaviour, state and non-state standard-setting, monitoring and behaviour modification processes [54].

Standard-setting can gain a strategic value for the weapon systems' cybersecurity. Therefore the strategy must include goals to:

- develop common definitions;
- create transparency;
- ensure compliance of the standards;

Standards for autonomous are designed and agreed on in cooperation with military and industry experts, lawyers, and academia with a technical, social and economic background. Standards can be open source and may also involve, in addition to European or US standardisation agencies, Asian or Russian agencies. As a shared interest and with an open standard transfer of knowledge and diffusion could become beneficial, the assumption is that this reduces risks and accidents. The exploitation of common standards could extend security dialogues and might reduce future military-to-military conflict. Via a dialogue, it is possible to open communications and gain a shared understanding of the autonomous weapon system evolution; it reduces the necessity to only focus on worst-case scenarios such as a battle-of-the-arms.

Bibliography

[1] United Nations, "Final Report of the 2019 Meeting of the High Contracting Parties to the CCW," 2019. [Online]. Available: https://www.unog.ch/80256EE600585943/(httpPages)/A0A0A3470E40345CC12580CD003D7927?OpenDocument.

[2] D. A. Lewis, G. Blum, and N. K. Modirzadeh, "WAR-ALGORITHM ACCOUNTABILITY," 2016. [Online]. Available: http://www.shirky.com/.

[3] P. Theron and A. Kott, "When Autonomous Intelligent Goodware Will Fight Autonomous Intelligent Malware: A Possible Future of Cyber Defense," in *Proceedings - IEEE Military Communications Conference MILCOM*, 2019, vol. 2019-Novem, doi: 10.1109/MILCOM47813.2019.9021038.

[4] US-GAO, "DOD Just Beginning to Grapple with Scale of Vulnerabilities," no. October, 2018, [Online]. Available: https://www.gao.gov/assets/700/694913.pdf.

[5] J. van den Berg *et al.*, "On (the emergence of) cyber security science and its challenges for cyber security education," in *The NATO IST-122 Cyber Security Science and Engineering Symposium.*, 2014, no. c, pp. 1–10.

[6] T. Rid, "Cyber War Will Not Take Place," *Journal of Strategic Studies*, vol. 35, no. 1, pp. 5–32, 2012, doi: 10.1080/01402390.2011.608939.

[7] T. Rid and B. Buchanan, "Attributing Cyber Attacks," *Journal of Strategic Studies*, vol. 38, no. 1–2, pp. 4–37, 2014, doi: 10.1080/01402390.2014.977382.

[8] R. Thornton and M. Miron, "Towards the 'Third Revolution in Military Affairs': The Russian Military's Use of AI-Enabled Cyber Warfare," *RUSI Journal*, vol. 165, no. 3, pp. 12–21, 2020, doi: 10.1080/03071847.2020.1765514.

[9] A. Elliott, "Automated mobilities: From weaponized drones to killer bots," *Journal of Sociology*, vol. 55, no. 1, pp. 20–36, 2019, doi: 10.1177/1440783318811777.

[10] B. Buruk, P. E. Ekmekci, and B. Arda, "A critical perspective on guidelines for responsible and trustworthy artificial intelligence," *Medicine, Health Care and Philosophy*, vol. 23, no. 3, pp. 387–399, Sep. 2020, doi: 10.1007/s11019-020-09948-1.

[11] M. Durante, "What Is the Model of Trust for Multi-agent Systems? Whether or Not E-Trust Applies to Autonomous Agents," *Knowledge, Technology & Policy*, vol. 23, no. 3–4, pp. 347–366, 2010, doi: 10.1007/s12130-010-9118-4.

[12] P. Timmers, "Strategic Autonomy and Cybersecurity," 2019. [Online]. Available: https://eucyberdirect.eu/content_research/strategic-autonomy-and-cybersecurity/.

[13] T. Everitt, "Towards Safe Artificial General Intelligence," 2018, [Online]. Available: http://hdl.handle.net/1885/164227.

[14] L. D. Introna, "Algorithms, Governance, and Governmentality: On Governing Academic Writing," *Science, Technology, & Human Values*, vol. 41, no. 1, pp. 17–49, 2016, doi: https://doi.org/10.1177/0162243915587360.

[15] R. v. Yampolskiy, *Artificial Superintelligence*. 2015.

[16] S. Ziesche and R. Yampolskiy, “Towards AI Welfare Science and Policies,” *Big Data and Cognitive Computing*, vol. 3, no. 1, p. 2, 2018, doi: 10.3390/bdcc3010002.

[17] R. v. Yampolskiy and M. S. Spellchecker, “Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures,” 2016, [Online]. Available: http://arxiv.org/abs/1610.07997.

[18] R. v Yampolskiy, *Artificial Intelligence Safety and Security*, First. Baco Raton: CRC Press, 2018.

[19] F. Pistono and R. v Yampolskiy, “Unethical Research: How to Create a Malevolent Artificial Intelligence,” pp. 1–7, 2016.

[20] P. Scharre, “Autonomous Weapons and Operational Risk Ethical Autonomy Project,” 2016.

[21] P. Scharre and M. C. Horowitz, “AUTONOMY in WEAPON SYSTEMS,” 2015. [Online]. Available: www.stopkillerrobots.org/.

[22] J. O. Birkeland, “The Concept of Autonomy and the Changing Character of War,” *Oslo Law Review*, vol. 5, no. 02, pp. 73–88, 2018, doi: 10.18261/issn.2387-3299-2018-02-02.

[23] P. Scharre, *Army of None: Autonomous Weapons and the Future of War*. W. W. Norton, 2018.

[24] B. M. Jensen, C. Whyte, and S. Cuomo, “Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence,” *International Studies Review*, vol. 22, no. 3, pp. 526–550, 2020, doi: 10.1093/isr/viz025.

[25] I. Verdiesen, F. Santoni de Sio, and V. Dignum, “Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight,” *Minds and Machines*, no. 0123456789, 2020, doi: 10.1007/s11023-020-09532-9.

[26] C. H. Heinl, “Artificial (intelligent) agents and active cyber defence: Policy implications,” in *International Conference on Cyber Conflict, CYCON*, 2014, vol. 2014, pp. 53–66, doi: 10.1109/CYCON.2014.6916395.

[27] G. Bills, “LAWS unto themselves: Controlling the development and use of Lethal Autonomous Weapons Systems,” *George Washington Law Review*, vol. 83, no. 1, pp. 176–208, 2014.

[28] T. Wischmeyer and T. Rademacher, *Regulating artificial intelligence*. Springer International Publishing, 2019.

[29] R. Kiggins, *The political economy of robots: prospects for prosperity and peace in the automated 21st century*. Cham: Springer Nature, 2018.

[30] S. D. Baum, “On the promotion of safe and socially beneficial artificial intelligence,” *AI and Society*, vol. 32, no. 4, pp. 543–551, 2017, doi: 10.1007/s00146-016-0677-0.

[31] M. R. Endsley, “From Here to Autonomy: Lessons Learned from Human-Automation Research,” *Human Factors*, vol. 59, no. 1, pp. 5–27, 2017, doi: 10.1177/0018720816681350.

[32] V. Boulanin and M. Verbruggen, “MAPPING THE DEVELOPMENT OF AUTONOMY IN WEAPON SYSTEMS,” 2017.

[33] F. Jahan, W. Sun, Q. Niyaz, and M. Alam, “Security modeling of autonomous systems: A survey,” *ACM Computing Surveys*, vol. 52, no. 5, 2019, doi: 10.1145/3337791.

[34] M. Dwyer, “Prioritizing Weapon System Cybersecurity in a Post-Pandemic Defense Department,” *Prioritizing Weapon System Cybersecurity in a Post-Pandemic Defense Department*, 2020. https://www.csis.org/analysis/prioritizing-weapon-system-cybersecurity-post-pandemic-defense-department (accessed Jan. 10, 2021).

[35] P. Mcdaniel, “ARTIFICIAL INTELLIGENCE AND CYBERSECURITY: OPPORTUNITIES AND CHALLENGES TECHNICAL WORKSHOP SUMMARY REPORT The National Science Technology Council’s NITRD and MLAI Subcommittees gratefully acknowledge,” 2020. [Online]. Available: http://www.whitehouse.gov/ostp.

[36] M. Zahid, I. Inayat, M. Daneva, and Z. Mehmood, “A security risk mitigation framework for cyber physical systems,” *Journal of Software: Evolution and Process*, vol. 32, no. 2, pp. 1–15, 2020, doi: 10.1002/smr.2219.

[37] G. McGraw, “Cyber War is Inevitable (Unless We Build Security In),” *Journal of Strategic Studies*, vol. 36, no. 1, pp. 109–119, 2013, doi: 10.1080/01402390.2012.742013.

[38] T. Stevens, “CYBERWEAPONS: POWER AND THE GOVERNANCE OF THE INVISIBLE,” *International Politics*, 2018.

[39] B. Jensen, B. Valeriano, and R. Maness, “Fancy bears and digital trolls: Cyber strategy with a Russian twist,” *Journal of Strategic Studies*, vol. 42, no. 2, pp. 212–234, Feb. 2019, doi: 10.1080/01402390.2018.1559152.

[40] K. D. Young, “The Militarization of Artificial Intelligence: A Systematic Inquiry Into the Reliability, Vulnerability, and Responsibility of Autonomous Weapons Systems,” *ProQuest Dissertations and Theses*, no. May, p. 78, 2017, [Online]. Available: https://search.proquest.com/docview/1897588762?accountid=10785%0Ahttp://linksource.ebsco.com/linking.aspx?sid=ProQuest+Dissertations+%26+Theses+Global&fmt=dissertation&genre=dissertations+%26+theses&issn=&volume=&issue=&date=2017-01-01&spage=&title=The+Mi.

[41] F. E. Morgan *et al.*, “Military applications of artificial intelligence : ethical concerns in an uncertain world,” Rand Corporation, Santa Monica, 2020.

[42] G. McGraw, R. Bonett, H. Figueroa, and V. Shepardson, “Security engineering for machine learning,” *Computer*, vol. 52, no. 8, pp. 54–57, Aug. 2019, doi: 10.1109/MC.2019.2909955.

[43] S. Cha, S. Baek, S. Kang, and S. Kim, “Security Evaluation Framework for Military IoT Devices,” *Security and Communication Networks*, vol. 2018, 2018, doi: 10.1155/2018/6135845.

[44] S. G. Brooks *et al.*, *The Rise and Fall of the Great Powers in the Twenty-first Century*, vol. 24, no. 1. 2016.

[45] R. Koch and M. Golling, “Weapons systems and cyber security-a challenging union,” in *International Conference on Cyber Conflict, CYCON*, 2016, vol. 2016-Augus, pp. 191–203, doi: 10.1109/CYCON.2016.7529435.

[46] J. Ferris, “Netcentric warfare, C4ISR and information operations: Towards a revolution in military intelligence?,” *Intelligence and National Security*, vol. 19, no. 2. pp. 199–225, 2004, doi: 10.1080/0268452042000302967.

[47] T. Moon, “Net-centric or networked military operations?,” *Defense and Security Analysis*, vol. 23, no. 1, pp. 55–67, Mar. 2007, doi: 10.1080/14751790701254474.

[48] R. Koch and M. Golling, “Weapons systems and cyber security-a challenging union,” in *International Conference on Cyber Conflict, CYCON*, 2016, vol. 2016-Augus, pp. 191–203, doi: 10.1109/CYCON.2016.7529435.

[49] M. Tarokh, M. Cross, and M. Lee, “Fuzzy logic decision making for multi-robot security systems,” *Artificial Intelligence Review*, vol. 34, no. 2, pp. 177–194, 2010, doi: 10.1007/s10462-010-9168-8.

[50] M. Dillon, “Virtual security: A life science of (dis)order,” *Millennium: Journal of International Studies*, vol. 32, no. 3, pp. 531–558, 2003, doi: 10.1177/03058298030320030901.

[51] M. M. Maas, “How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons,” *Contemporary Security Policy*, vol. 40, no. 3, pp. 285–311, Jul. 2019, doi: 10.1080/13523260.2019.1576464.

[52] L. Armand, “The posthuman: AI, dronology, and ‘becoming alien,’” *AI and Society*, vol. 35, no. 1, pp. 257–262, 2020, doi: 10.1007/s00146-018-0872-2.

[53] L. Lessig, *Code: version 2.0*. Basic Books, 2006.

[54] B. Morgan and K. Yeung, *An introduction to law and regulation: Text and materials*. 2007.

[55] M. Dunn Cavelty, “From cyber-bombs to political fallout: Threat representations with an impact in the cyber-security discourse,” *International Studies Review*, vol. 15, no. 1, pp. 105–122, 2013, doi: 10.1111/misr.12023.

[56] M. Dunn Cavelty, “The militarisation of cyberspace: Why less may be better,” 2012.

[57] N. Davison, “Autonomous weapon systems: An ethical basis for human control? - Humanitarian Law & Policy Blog | Humanitarian Law & Policy Blog,” 2018. https://blogs.icrc.org/law-and-policy/2018/04/03/autonomous-weapon-systems-ethical-basis-human-control/.

[58] M. Verbruggen, “The Role of Civilian Innovation in the Development of Lethal Autonomous Weapon Systems,” *Global Policy*, vol. 10, no. 3, pp. 338–342, Sep. 2019, doi: 10.1111/1758-5899.12663.

[59] “A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977: International Committee of the Red Cross Geneva, January 2006,” *International Review of the Red Cross*, vol. 88, no. 864. International Committee of the Red Cross, Geneva, pp. 931–956, 2006, doi: 10.1017/S1816383107000938.

[60] H. Y. Liu, “Categorization and legality of autonomous and remote weapons systems,” *International Review of the Red Cross*, vol. 94, no. 886, pp. 627–652, 2013, doi: 10.1017/S181638311300012X.

[61] M. A. C. Ekelhof, “Complications of a common language: Why it is so hard to talk about autonomous weapons,” *Journal of Conflict and Security Law*, vol. 22, no. 2, pp. 311–331, Jun. 2017, doi: 10.1093/jcsl/krw029.

[62] B. M. J. Griffor Edward R., Greer Christopher, Wollman David A., “Framework for Cyber-Physical Systems : Volume 1 , Overview NIST Special Publication 1500-201 Framework for

Cyber-Physical Systems : Volume 1 , Overview," *Nist*, vol. 1, no. 1, p. 79, 2017, [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-201.pdf.

[63] H. A. Kholidy, "Autonomous mitigation of cyber risks in the Cyber–Physical Systems," *Future Generation Computer Systems*, vol. 115, pp. 171–187, Feb. 2021, doi: 10.1016/j.future.2020.09.002.

[64] K. Anderson and M. C. Waxman, "Debating Autonomous Weapon Systems, Their Ethics, and Their Regulation Under International Law," *The Oxford Handbook of Law, Regulation, and Technology*, vol. 553, no. 14, pp. 1097–1117, Feb. 2017, doi: 10.1093/oxfordhb/9780199680832.013.33.

[65] K. Anderson and M. C. Waxman, "Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can," *SSRN Electronic Journal*, 2013, doi: 10.2139/ssrn.2250126.

[66] C. P. Trumbull, "Autonomous Weapons: How Existing Law Can Regulate Future Weapons," *SSRN Electronic Journal*, vol. 321, 2019, doi: 10.2139/ssrn.3440981.

[67] UNIDIR, "Framing Discussions on the Weaponization of Increasingly Autonomous Technologies," *UNIDIR Resources*, no. 1, 2014, [Online]. Available: http://www.unidir.org/files/publications/pdfs/framing-discussions-on-the-weaponization-of-increasingly-autonomous-technologies-en-606.pdf.

[68] H. M. Roff, "Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous Weapons," *Global Security Initiative*, pp. 1–8, 2016, [Online]. Available: http://www.article36.org/wp-content/uploads/2016/12/Control-or-Judgment_-Understanding-the-Scope.pdf.

[69] G. J. Schaub Jr and J. W. Kristoffersen, "In, On, or Out of the Loop?: Denmark and Autonomous Weapon Systems," 2017, [Online]. Available: http://cms.polsci.ku.dk/.

[70] D. Snyder, J. D. Powers, E. Bodine-Baron, B. Fox, L. Kendrick, and M. H. Powell, *Improving the Cybersecurity of U.S. Air Force Military Systems Throughout Their Life Cycles*. 2015.

[71] UNIDIR, "Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies," *UNIDIR Resources*, vol. 5, no. 5, 2016, [Online]. Available: www.unidir.org.

[72] M. Lodge, "Regulation, the regulatory state and European politics," *West European Politics*, vol. 31, no. 1–2, pp. 280–301, 2008, doi: 10.1080/01402380701835074.

[73] A. Dafoe, "AI Governance: A Research Agenda," 2018. Accessed: Jun. 18, 2020. [Online]. Available: www.fhi.ox.ac.uk/govaiagenda.

[74] C. Ai, "Limitations on current AI technology," *The Foundations of Artificial Intelligence*, pp. 381–382, 2010, doi: 10.1017/cbo9780511663116.037.

[75] Y. Feng, M. Li, C. Zeng, and H. Liu, "Robustness of internet of battlefield things (IoBT): A directed network perspective," *Entropy*, vol. 22, no. 10, pp. 1–15, Oct. 2020, doi: 10.3390/e22101166.

[76] B. Valeriano and A. Craig, “Reacting to Cyber Threats: Protection and Security in the Digital Age,” *Global Security and Intelligence Studies*, vol. 1, no. 2, pp. 21–41, 2016, doi: 10.18278/gsis.1.2.2.

[77] T. Stevens, “Global Cybersecurity: New Directions in Theory and Methods,” *Politics and Governance*, vol. 6, no. 2, p. 1, Jun. 2018, doi: 10.17645/pag.v6i2.1569.

[78] I. Nai-Fovino, R. Neisse, J. L. Hernandez-Ramos, N. Polemi, G. Ruzzante, and M. Figwer, “A Proposal for a European Cybersecurity Taxonomy,” 2019, doi: 10.2760/106002.

[79] P. Ducheine and Peter. Pijpers, “The notion of cyber operations,” *Amsterdam Law School Research Paper*, vol. 202, no. 08, pp. 211–232, 2020, doi: http://dx.doi.org/10.2139/ssrn.3575755.

[80] P. Ducheine and J. van Haaster, “Fighting power, targeting and cyber operations,” *International Conference on Cyber Conflict, CYCON*, vol. 2014, pp. 303–327, 2014, doi: 10.1109/CYCON.2014.6916410.

[81] P. W. Singer and A. Friedman, *Cybersecurity: What Everyone Needs to Know*. OUP USA, 2014.

[82] United States: US Army, “Cyberspace Operations Concept Capability Plan 2016-2028,” *TRADOC Pamphlet 525-7-8*, no. February 2010, pp. 1–77, 2010, [Online]. Available: https://fas.org/irp/doddir/army/pam525-7-8.pdf%0Ahttp://www.fas.org/irp/doddir/army/pam525-7-8.pdf.

[83] B. Valeriano and R. C. Maness, *Cyber War versus Cyber Realities: Cyber Conflict in the International System*. Oxford University Press, 2015.

[84] J. van den Berg, “Nationale veiligheid en crisisbeheersing,” *Nationale veiligheid en crisisbeheersing*, vol. 13, no. 2, pp. 4–5, 2015, Accessed: Feb. 11, 2019. [Online]. Available: https://www.ifv.nl/kennisplein/Documents/20150511-NCTV-Magazine-nationale-veiligheid-en-crisisbeheersing.pdf.

[85] R. J. Deibert and R. Rohozinski, “Risking security: Policies and paradoxes of cyberspace security,” *International Political Sociology*, vol. 4, no. 1, pp. 15–32, Mar. 2010, doi: 10.1111/j.1749-5687.2009.00088.x.

[86] J. Lewis, “Cybersecurity and critical infrastructure protection,” 2006. doi: 10.1016/b978-0-12-817137-0.00008-0.

[87] D. Craigen, N. Diakun-Thibault, and R. Purse, “Defining Cybersecurity,” *Technology Innovation Management Review*, vol. 4, no. 10, pp. 13–21, 2014, doi: 10.22215/timreview835.

[88] T. Stevens, *Cyber security and the politics of time*. 2015.

[89] B. van den Berg and E. Keymolen, “Regulating security on the Internet: control versus trust Regulating security on the Internet: control versus trust,” 2017, doi: 10.1080/13600869.2017.1298504.

[90] S. O. Hansson, “Seven Myths of Risk,” 2005. Accessed: Jun. 05, 2019. [Online]. Available: https://graelaws.files.wordpress.com/2011/01/risk-managemntan-international-journal2.pdf.

[91] L. Zedner, *Security*. New York: Taylor & Francis, 2009.

[92] NIST, “NIST Special Publication 800-30 Revision 1 - Guide for Conducting Risk Assessments,” *NIST Special Publication*, no. September, p. 95, 2012, doi: 10.6028/NIST.SP.800-30r1.

[93] O. Renn, “Concepts of risk: a classification,” in *Social Theories of Risk*, no. January 1992, 1992, pp. 53–79.

[94] C. C. Demchak, “China: Determined to dominate cyberspace and ai,” *Bulletin of the Atomic Scientists*, vol. 75, no. 3, pp. 99–104, May 2019, doi: 10.1080/00963402.2019.1604857.

[95] C. Demchak, “Cybered conflict, cyber power, and security resilience as strategy,” in *Cyberspace and National Security: Threats, Opportunities, and Power in a Virtual World*, 2012, pp. 121–136.

[96] S. L. Pfleeger, “Leveraging Behavioral Science to Mitigate Cyber Security Risk Leveraging Behavioral Science to Mitigate Cyber Security Risk © 2012 The MITRE Corporation . ALL RIGHTS RESERVED .,” *Conputers and Security*, pp. 1–44, 2012, doi: 10.1016/j.cose.2011.12.010.

[97] I. S. O. ISO, “IEC 27005: Information technology–security techniques–information security risk management,” *Iso/Iec*, vol. 44, 2011, [Online]. Available: www.iso.org.

[98] N. C. S. D. CSD, “NIST SP 800-53 Revision 3, Recommended Security Controls for Federal Information Systems and Organizations,” pp. 1–237, 2014, [Online]. Available: papers3://publication/uuid/8ED3C786-E3B5-4C99-BCBB-EA60823492AE.

[99] Q. J. Yeh and A. J. T. Chang, “Threats and countermeasures for information system security: A cross-industry study,” *Information and Management*, vol. 44, no. 5, pp. 480–491, 2007, doi: 10.1016/j.im.2007.05.003.

[100] W. Jansen and T. Grance, “NIST Special Publication 800-144 Guidelines on Security and Privacy in Public Cloud Computing,” *Nist Special Publication*, vol. 144, no. 7, pp. 1–70, 2011, Accessed: May 13, 2019. [Online]. Available: https://cloudsecurityalliance.org/download/the-treacherous-twelve-.

[101] E. Ntoutsi *et al.*, “Bias in data-driven artificial intelligence systems—An introductory survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, pp. 1–14, 2020, doi: 10.1002/widm.1356.

[102] U. Lindqvist and E. Jonsson, “How to systematically classify computer security intrusions,” *IEEE Symposium on Security and Privacy (Cat. No. 97CB36097)*, no. 1530, pp. 154–163, 1997, doi: 10.1109/secpri.1997.601330.

[103] B. Carter *et al.*, “A Preliminary Design-Phase Security Methodology for Cyber–Physical Systems,” *Systems*, vol. 7, no. 2, p. 21, 2019, doi: 10.3390/systems7020021.

[104] W. Young and N. Leveson, “Systems thinking for safety and security,” *ACM International Conference Proceeding Series*, pp. 1–8, 2013, doi: 10.1145/2523649.2530277.

[105] G. J. Nalepa, M. Araszkiewicz, S. Nowaczyk, and S. Bobek, “Building trust to AI systems through explainability. Technical and legal perspectives,” in *CEUR Workshop Proceedings*, 2019, vol. 2681.

[106] M. Lodge and Kai. Wegrich, *Managing Regulation: Regulatory Analysis, Politics and Policy*. Basingstoke: Palgrave Macmillan, 2012.

[107] Myriam. Dunn Cavelty, Victor. Mauer, and S. Felicia. Krishna-Hensel, *Power and security in the information age: Investigating the role of the state in cyberspace*. Ashgate, 2008.

[108] S. Fugate and K. Ferguson-Walter, “Artificial intelligence and game theory models for defending critical networks with cyber deception,” *AI Magazine*, vol. 40, no. 1, pp. 49–62, 2019, doi: 10.1609/aimag.v40i1.2849.

[109] M. Klincewicz, “Autonomous Weapons Systems, the Frame Problem and Computer Security,” *Journal of Military Ethics*, vol. 14, no. 2, pp. 162–176, 2015, doi: 10.1080/15027570.2015.1069013.

[110] Y. Qian, Y. Fang, and J. J. Gonzalez, “Managing information security risks during new technology adoption,” *Computers and Security*, vol. 31, no. 8, pp. 859–869, 2012, doi: 10.1016/j.cose.2012.09.001.

[111] S. Duindam, “The public good ‘Defense,’” in *Military Conscription*, W. Muller and M. Bihn, Eds. Springer Berlin Heidelberg, 1999.

[112] R. Zwetsloot and A. Dafoe, “Thinking About Risks From AI: Accidents, Misuse and Structure,” 2019. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure (accessed Dec. 12, 2020).

[113] J. B. Wiener, “The regulation of technology, and the technology of regulation,” *Technology in Society*, vol. 26, no. 2–3, pp. 483–500, Apr. 2004, doi: 10.1016/j.techsoc.2004.01.033.

[114] P. Cihon *et al.*, “Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development,” 2019. [Online]. Available: https://arxiv.org/pdf/1802.07228.pdf.

[115] K. Blind, “The impact of standardisation and standards on innovation,” in *Handbook of Innovation Policy Impact*, J. Edler, P. Cunningham, and A. Gök, Eds. Cheltenham: Edward Elgar Publishing, 2016, pp. 423–450.

[116] T. Stevens, “Cyberweapons: an emerging global governance architecture,” *Palgrave Communications*, vol. 3, no. 1, pp. 1–6, 2017, doi: 10.1057/palcomms.2016.102.

Appendix 1

Semi Structured Interview Protocol

- What is your view of autonomous weapons systems? Are you enthusiastic or worried about these technological developments?
- What is your view of Artificial Intelligence when it is going to be used to increase the (cyber) security of weapon systems?
- What do you see as the most relevant features or techniques? Do you see a role for you of AI in protecting AWS? What role do you see in securing AWS?
- What risks / worries / possibilities do you see or experience when using these technologies?
- Trust is seen as a prerequisite for AI and thus AWS. How can security help to maintain trust in AWS?
- Transparency is seen as another principle that is needed, what is your vision of transparency in the case of AWS?
- What steps should we take to ensure responsible use of AI in AWS?
- What principles, standards or frameworks are needed in the case of military use, what differences do you see between civilian and military use?
- How do you think the defense of AWS can be regulated?
- What can regulations in the form of standards, principles or laws mean for the development of AWS ?