University of Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/Specialization: RISK ANALYSIS AND GOVERNANCE	Spring semester, 2021 Open / ~~Restricted~~ access
Writer: TOLUWABORI MAGBAGBEOLA	………………………………………… (Writer's signature)

Faculty supervisor: REIDAR STAUPE-DELGADO

External supervisor(s):

Thesis title:

ASSESSING AND MANAGING THE RISK OF USING LETHAL AUTONOMOUS WEAPONS

Credits (ECTS): 30 ECTS

Key words: Lethal Autonomous Weapons Systems, Risk, System Failure, Normal Accident Theory, High Reliability Theory, Safety, Risk Regulations, Weapons Review, Proportionality, Distinction, Precautionary principle	Pages: 73 + enclosure: N/A Stavanger, 14th June 2021 Date/year

ABSTRACT

Over the last two (2) decades, there has been a surge in scholarly attention and a lot has been written on the lethal autonomous weapons systems (also known as “Killer Robot”). The focus of writing has been on the legal, ethical, moral, and policy issues pertaining to Lethal Autonomous Weapons System (LAWS). Thus, a lot of attention and concerns have been directed to what happens when a Lethal Autonomous Weapon System goes wrong? However, little or no attention has been directed to discussions such what are the risks surrounding the development, deployment, and use of Lethal Autonomous Weapons System LAWS? How should the international community including countries address risk regulations of the Lethal Autonomous Weapon System if there are uncertainties as to how these systems may fail?

This thesis addresses the risks attached to Lethal Autonomous Weapons System, complex, tightly coupled and unpredictable high-risk technology. As such, the thesis debates the risks to LAWS from the Normal Accident Theory perspective and the discusses the uncertainty/unpredictability revolving around LAWS. It goes further to argue whether the high-reliability theory can be used as means of safety in Lethal Autonomous Weapon Systems.

In addition, this thesis addresses the obstacles of LAWS complying with risk regulations. In discussing that, it argues whether LAWS without human intervention can be said to appear to be capable of complying with the key principles of risk regulations such as the international humanitarian law and the laws of armed conflicts principles of proportionality, distinction, and precaution to bring about societal safety to the community where LAWS is deployed or engaged.

This thesis is expected to direct attention and focus to discussions surrounding the risks of developing, engaging and deploying LAWS and the obstacles that LAWS presents in complying with risk regulations that can bring about safety in the use and deployment of LAWS.

Keywords: Lethal Autonomous Weapons Systems, Risk, System Failure, Normal Accident Theory, High Reliability Theory, Safety, Risk Regulations, Weapons Review, Proportionality, Distinction, Precautionary principle

PREFACE

This Master's thesis is written and submitted as the final part of an M.Sc. in Risk Analysis and Governance under the administration of the Faculty of Science and Technology, University of Stavanger, Norway, and the supervision of Professor Reidar Staupe-Delgado of the University of Stavanger.

The thesis is a partial fulfillment of the requirements needed for being awarded a Master of Science degree.

I would like to thank Reidar Staupe-Delgado for his guidance and support in helping me complete this thesis.

ABBREVIATIONS

A.I – Artificial Intelligence

A.N.N - Artificial Neural Network

A.M.R.A.A.M – Advanced Medium Range Air to Air Missile

CCW – Convention on Certain Conventional Weapons

DoD – Department of Defense

HRT – High Reliability Theory

I.C.J – International Court of Justice

IFF- Identification Friend or Foe

IHL – International Humanitarian Law

LAWS – Lethal Autonomous Weapons Systems

LOAC – Law of Armed Conflict

M.L. – Machine Learning

NASA – National Aeronautics and Space Administration

NATO – The North Atlantic Treaty Organization

NDAA – National Defense Authorization Act

NLP – National Language Processing

SUBSAFE – Submarine Safety Program

UN- United Nations

UNIDR – United Nations Institute for Disarmament Research

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The growth of AI research and its application has become parallel to the world's industrialization and economic development, and it has affected every facet of our lives. Machines with varying levels of intelligence now exist to perform basic and some not so basic human tasks but increasingly complex operations in various fields of endeavours such as medicine, health care, logistics, transportation, militaries, etc. (Brynjolfsson & McAfee, 2014). At each facet of our lives, AI can be recognized as a powerful and incredible tool and yet a tool that can endanger lives. The application of AI in the weapons industry has become a subject of interest and a cause for debate in recent years. Lethal Autonomous Weapon Systems, popularly referred to by some as "killer robots", has been a subject of consistent discussions within the framework of the Convention on Certain Conventional Weapons (CCW). (1342 UNTS 137) and will continue to be debated in the years to come. And such discussions are going to fundamentally change and influence the way wars and battles are waged in the future. (Singer, 2012).

Lethal Autonomous Weapons Systems (LAWS) are a unique set of systems with a clear distinction from any forms of prior weaponry. LAWS' ability to self-decide by selecting and engaging targets without human intervention or oversight makes them extremely effective and unpredictable. (Crootof, 2014). Besides, LAWS have raised questions that are not only technical or military but also legal, ethical, and socio-political in nature. There are different views on whether the development and deployment of LAWS are ethical, moral, legal, or desirable by the international community. To answer these questions, two schools of thought exist.

The first school of thought emphasizes the benefits and importance of LAWS. (Schmitt, 2013). Autonomous systems can easily identify and process complex information at an extremely high speed; make faster and more precise decisions; save lives by replacing human combats and reducing the number of casualties involved; this specific school of thought has argued that due to lack of emotions such as fear, revenge, frustration, anger, selfishness, fatigue, hysteria, etc., LAWS can make objective analysis and decisions without bias or emotions to cloud its judgment. With LAWS, atrocities and killings like the My Lai massacre or Fallujah will not exist (although this is subjective and debatable).

The second school of thought believes and views Lethal Autonomous Weapons System as a threat to the international community. According to this school of thought, there will be no value on human lives once a machine is given the power to make decisions to kill. LAWS is believed to be the beginning of the arms race, the death of us all, and LAWS will be abused and misused by society. Due to the complexity of the component systems and subsystems, it is believed that LAWS will experience several malfunctions and system failures, battles will become disproportionate and sometimes rigid because LAWS have no emotion or compassion to analyze the situations and environments of the battlefield before making decisions, they are isolated from the battle and their targets, and indifferent towards the outcome of their decisions and actions. For LAWS, the battle becomes a line of command, unreal, and humans become data, enemies become inhuman and irrelevant.

Over the last two (2) decades, there has been a surge in scholarly attention and a lot has been written on the lethal autonomous weapons systems (also known as “Killer Robot”). The focus of writing has been on the legal, ethical, moral, and policy issues pertaining to LAWS. Thus, a lot of attention and concerns have been directed to what happens when a Lethal Autonomous Weapon System goes wrong? However, little or no attention has been directed to discussions such as how will Lethal Autonomous Weapon System go wrong? What are the risks surrounding LAWS? How should the international community including countries address risk regulations of the Lethal Autonomous Weapon System if there are uncertainties as to how these systems may fail? Can an ironclad risk regulation bring about societal security and safety?

In giving insights into LAWS, one of the most important things to bear in mind is that LAWS is different from the prior operational weapon systems. With AI, LAWS can make decisions on its own without human intervention and this creates a huge gap in understanding how LAWS can go wrong and also the unique challenges in regulating LAWS from a risk perspective. The debates on how to deal with these problems have become a matter of concern to the international community.

This thesis fills that gap by using the Normal Accident Theory proposed by Charles Perrow to address the concerns of how LAWS can go wrong through the system failure lens. Although Charles Perrow applied his theories to high-risk systems such as nuclear power plants, aircraft/airspace, and marines, authors such as Sagan (1995) and Snook (2000) have expanded

Charles Perrow's Normal Accident theory to other sectors such as nuclear weapons and military system, respectively. Lethal Autonomous Weapons System is a high-risk system that falls under the military classification even though private individuals/companies in Silicon Valley are currently in charge of the development of LAWS.

This paper takes a step further to criticize the normal accident theory and view LAWS from the High Reliability Theory which argues that "extremely safe operations are possible even with extremely hazardous technologies if appropriate organizational design and management techniques are followed" (Sagan, 1995 p. 13). In addition, this thesis highlights the possibility of LAWS being incapable of complying with key risk regulation due to lack of human intervention and addresses the incapability of LAWS to ever adequately make highly complex and contextual analyses that international humanitarian law (IHL) requires. This thesis finally highlights some of the key principles of IHL that policymakers need to consider for risk regulations for Lethal Autonomous Weapons.

To properly address these gaps, this thesis shall answer the following analytical questions:

i. What are the risks attached to complex, tightly coupled, and unpredictable high-risk technology like Lethal Autonomous Weapons?
ii. What safety model exists for complex, tightly coupled, and unpredictable high-risk technology like Lethal Autonomous Weapons?
iii. What is the security perspective of LAWS?
iv. Can LAWS comply with risk regulation to bring about societal safety and security?

In answering these analytical questions, this paper is divided into five parts. The first part (Chapter 2) will define several key theoretical frameworks, diverging views, and review works of literature that were considered in writing this thesis. The second part (Chapter 3) considers the underlying technology of Lethal Autonomous Weapon System –AI and extensively discusses the Lethal Autonomous Weapon Systems. The third part (Chapter 4) of this thesis debates the risk of Lethal Autonomous Weapons from the Normal Accident Theory perspective and the uncertainty/unpredictability revolving around LAWS. It goes further to argue whether the high-reliability theory can be used as means of safety in Lethal Autonomous Weapon Systems. The fourth part (Chapter 5) argues whether lethal autonomous weapons systems can be said to appear to be capable with the help of algorithm to abide by key principles of risk regulations such as the international humanitarian law and the laws of armed conflicts principles of proportionality, distinction, and precaution which call for unquantifiable decision-making. Finally, the fifth part (Chapter 6) holds an intellectual discussion on the earlier discourse and offers solutions and recommendations to Lethal Autonomous Weapon Systems.

LAWS are not reinventing the wheel; they are wheels that can spin themselves. Even though it is highly unlikely that countries and the international community will deploy LAWS that are unpredictable or non-compliant with the risk regulation and principles of international humanitarian law if such is deployed nevertheless, the consequences of such may be severe on the entire international community. Thus, this research thesis shall demonstrate how Lethal Autonomous System is spinning the wheel and I expect this thesis to contribute to debates revolving around risks of the Lethal Autonomous Weapons System and the compliance of LAWS with risk regulations that have the potential of reducing some of the risks of LAWS.

When citing LAWS, this thesis refers to fully autonomous weapons systems, Occasionally, the term autonomous weapon system (AWS), lethal autonomous systems, lethal autonomous weapons are used interchangeably in this research thesis.

In writing this thesis, qualitative research was chosen over quantitative research because the information and data available to write this thesis were from a theoretical standpoint with no clear association or interaction with numerical data. In using the qualitative research method, historical analysis was included to better understand the Lethal Autonomous Weapon Systems and provide the necessary direction of how to address the normal accident risk of LAWS.

Conceptual methodology of research was also used to shift the attention to and provide further information relating to the risk of Lethal Autonomous Weapons System and re-interpret the existing information on the risk of LAWS in an understandable manner.

In narrowing the focus on the qualitative method of research used, content analysis was chosen as the approach to be used to answer the problem statements. With content analysis, this thesis was able to collate documentary materials that serve as evidence for the risk behaviour that is being assessed in Lethal Autonomous Weapons System and the contents of the documents were analyzed to form a deductive logical conclusion that was used in answering the problem statements (analytical questions) and bridging the gap between lethal autonomous weapons system, the development and use from the risk perspective.

CHAPTER 2

LITERATURE REVIEW: CONSIDERING DIVERGING PERSPECTIVES ON THE ISSUE

When it comes to debates and discussions on Lethal Autonomous Weapons System, two (2) different schools of thought have their views on the use of the Lethal Autonomous Weapons System. The first school of thought believes that the development and use of LAWS should be banned due to the inherent danger of using lethal autonomous weapons and the legal and ethical issues surrounding it. The second school of thought believes that Lethal Autonomous Weapon System is the messiah due to its reliability and precision, and it is the solution to some of the germane problems of warfare. In fact, the second school of thought believes that lethal autonomous weapons will change the perspective of future warfare.

These conflicting schools of thought do have a basis, in reality, to be right, and in many situations, autonomous systems have demonstrated the above-mentioned characteristics. For many people, as long as LAWS can adequately perform their task with precision, under normal operating conditions they may very well be better than humans. However, their brittle nature means that if LAWS is pushed beyond the normal bounds of their programming, they may fail miserably.

United States, Russia, China, and some other developed countries including Norway are investing heavily in the development of lethal autonomous weapon systems.

The question on every curious mind is: What is the driving force behind the lethal autonomous weapons system? To answer that question, two forces are driving these weapons:

- Technology: AI technology is the underlying technology for LAWS and it is advancing exponentially. It is the core of its existence and provides the intelligence required for lethal autonomous weapon systems (LAWS).
- Humanity: In 2016, the World Economic Forum (WEF) attendees were asked, “If your country was suddenly at war, would you rather be defended by the sons and daughters of your community or an autonomous AI weapons system?” (Marijan, 2016). The majority responded that they would choose Autonomous AI Weapons Systems over the sons and daughters of the community. This suggests that there is a common desire to have

autonomous weapons/robots fight rather than sending humans to risk their lives on the battlefield.

Autonomous weapon systems have raised a host of questions and the attention has been on the legality, morality, ethicality, responsibility gap, accountability, humanitarian and policy issues which have led some academics and industries to call for the ban of lethal autonomous weapons also known as "killer robot" at the international level. The focus and attention on the above-mentioned area have created a gap in addressing the risk of Lethal Autonomous Weapons Systems due to the complexity and tight coupling of the underlying technology of LAWS which is known as AI. There has been a lack of attention on the scope of system failures leading to accidents and how these failures can occur due to the development and deployment of Lethal Autonomous Weapons Systems.

Understanding the risks associated with the development and use of LAWS and addressing the gap that has been ignored by the international community is quite important for policymakers, innovators, developers, and controllers in deciding the development, use, and deployment of lethal autonomous weapons. To understand properly, it is important to look into the following theoretical framework:

2.1. Normal Accident Theory

When it comes to high-risk technologies such as Lethal Autonomous Weapons, no matter the control measures put in place or the resilience of the system, normal accidents are bound to occur due to the complexity of interaction and tight coupling of the system. According to Charles Perrow, *"if interactive complexity and tight coupling system characteristics inevitably will produce an accident, I believe we are justified in calling it a normal accident, or a system accident"*. (Perrow, 2011 p. 17-18). Even though it is infrequent, uncommon, and rare to have a system or normal accidents, when they occur, they can cause catastrophes and end the world.

In addressing the normal accidents that occur during the use of lethal autonomous weapons systems, it is quite important to account for various debates, schools of thought, and gaps that have been raised concerning normal accidents and lethal autonomous weapons.

According to Perrow, an accident can be regarded as an unintended and untoward event with uncertainties remaining. Therefore, *"an accident is a failure in a subsystem or the system as a whole, that damages more than one unit and in doing so disrupts the ongoing or future output of the system."* (Perrow, 2011 p. 90). Two types of accidents have been identified which are: component failure accidents and system accidents. With component failure accidents, there are one or more component failures in either the part, unit, or subsystems that are linked together in an expected sequence while system accident involves the interaction of multiple levels of failure that was unexpected or unanticipated.

Thus, the distinguishing factor is whether the interaction of one or multiple failures was expected, anticipated, or understandable to the person who designed and trained the system. Thus, the distinguishing factor is the occurrence or not of multiple failures interacting in an unexpected way.

According to Charles Perrow's theory of Normal Accident, he argues that the two system characteristics which are *"interactive complexity and tight coupling…inevitably will produce an accident"* in high-risk systems. (Perrow, 2011 p. 5). He also states that the unavoidable outcome of high-risk systems is accidents regardless of the safety measures taken, organization structure, or chain of command in place. His theory of Normal Accidents in "high-risk systems" has also been applied to nuclear weapons and military systems by authors such as Scott D. Sagan (1995) and Scott A. Snook (2000).

Flowing from the premise of Normal Accidents in high-risk systems, debates about lethal autonomous weapons can be framed and considered from the perspective of accidents and "high-risk technologies. Therefore, Lethal Autonomous Weapon Systems can be viewed as a tightly coupled and highly interactive complex system where the occurrence of accident could be as a result of the unanticipated interaction of the system components and it is inevitable regardless of standard operating systems and procedures, precautionary measures, safety mechanisms, and weapon reviews. When an accident eventually occurs, it is impossible to detect the reason why the failure occurred thereby making it extremely challenging to assign responsibilities for the failure.

2.1.1. System Interactions

The systems' parts, subsystems, or units interact with one another. The existence of such interaction does not pose any challenge provided that the interactions are not unanticipated and obvious with regards to the decision line.

Accident models have developed into various interactive phases such as simple linear models, complex linear models, and complex non-linear models. (Pryor & Capra, 2012). Simple and complex linear interactions are not the focus of this review although it is quite important to note that with linear interactions (whether simple or complex), when a system part fails, it becomes easily identifiable what happens to the system or system parts that have failed and the series and sequences of steps and decisions to be made. It is recognized as the interaction of one component in the systems with one or more components preceding it immediately in the sequence. (Perrow C., 2011 p. 106). It is designed to function and operate in this specific way thereby making it is easy to identify failure and know its effect.

2.1.2. Complex Non-Liner Interactions

Due to technological advancement, complex non-linear models were developed for systems and this complexity has led to inevitable and unavoidable system accidents that are now considered and recognized as normal accidents. (Pryor & Capra, 2012).

For non-linear interactions, the parts, subsystems, or units serve multiple functions thereby such interactions are unanticipated. The complexity of the interaction shows that the sequences are unplanned and unexpected. Thus, the complexity makes the failure difficult to identify because an unanticipated relationship between two or more subsystems, units, or parts may occur causing unexpected and unplanned interaction. At that point, the system operators have no clue of knowing that a failure has occurred until the outcome of the failure becomes evident. Therefore, making it difficult for operators and controllers to respond to such failures immediately because of their incomprehension. With complex interactions, one component interacts with one or more other components and the interactions are unfamiliar and unexpected sequences that are either not visible to the eyes or not comprehensible immediately. (Perrow, 2011).

According to Charles Perrow,

> *"……I will refer to these kinds of interactions as complex interactions, suggesting that there are branching paths, feedback loops, jumps from one linear sequence to another because of proximity……. The connections are not only adjacent, serial ones, but can multiply as other parts or units or subsystems are reached. The much more common interactions, the kind we intuitively try to construct because of their simplicity and comprehensibility, I will call linear interactions."* (Perrow, 2011 p. 104)

For complex systems, pulling out or shutting down a component of the system could lead to a temporary shut down or severance to numerous other components because parts and units tend to be linked in multiple directions. Those operating complex systems are less likely to recognize, predict, identify the interdependency failure before the escalation of the incident to an accident. (Perrow, 2011).

2.1.3. Tight Coupling

The existence of tight coupling in a system is a mechanism showing that there exists little or no slack or buffer between two subsystems thereby causing the actions affecting one subsystem directly to affect the other subsystem. (Perrow, 2011). Thus, interacting failures occurring in one system can move quickly without any obstruction whatsoever throughout the system thereby escalating the failures to system accidents. (Rijpma, 1997 p. 16).

According to Rijpma, it is common to find tight coupling among systems that are operating unfinal, invariant, time-dependent production processes, where safety devices are in-built and where improvisation is almost impossible (Rijpma, 1997). However, it should be noted that tight coupling is also common with high-risk technologies such as AI. When failures occur with tightly coupled systems, it is unlikely that such failure can be nipped in the bud, prevented at the point of occurrence, or quick recovery and restoration before the failure propagates swiftly because once the failure process begins, the system cannot be turned off.

Time dependence is one of the characteristics of a tightly coupled system. They cannot be left unattended because reactions within such systems are instant and cannot be extended or delayed. In addition, there is invariancy with the sequences that exist in tightly coupled systems. A lethal autonomous weapon must search, scan, observe, orient, decide and engage. The sequence cannot be swapped or rerouted to have the autonomous weapon orient, decide, scan, observe and then engage. Not only is the sequence invariant with tightly coupled systems, but the way the system is designed based on the variance also allows the system to reach its goals and engage its targets in a specific way. (Perrow, 2011 p. 127).

To prevent an incident and failure occurring in a subsystem from spreading and affecting other subsystems in a tightly coupled system, recovery from failures is important for consideration. Even though tightly coupled systems allow little or no slack or buffer; designers must be deliberate, and it must be carefully thought of in advance how safety devices or processes such as redundancies, substitutions, and buffers can be designed into the system.

An accident is inevitable for a complex and tightly coupled system. And such accidents are called Normal Accidents by Charles Perrow.

2.2. High-Reliability Theory as Safety Model?

Once a system is recognized as safety-critical, mechanisms should be put in place that will be useful in helping detect, prevent and/or tolerate a system failure that is known to be a normal accident. Implementing a model of safety ensures that normal accidents are avoided.

The High-Reliability Theory (HRT) contradicts Charles Perrow's Theory of Normal Accident. According to the High-Reliability Theory proponents, strategies that allow organizations with complex and tightly coupled systems to attain an outstanding level of safety were claimed to be discovered by the Berkley School on High-Reliability Theory. (Robert, 1993). Redundancy is highly used in organizations with high-reliability theory as a backup for failure that could either occur in systems or persons. (La Porte and Consolini, 1991).

Organizations that use high-reliability theory believe that they can avoid normal accidents even though it is expected to occur due to complexity and tight coupling. With a redundancy strategy, the existence of a backup in the event of the failure of a component system or subsystem reduces the probability of the occurrence of failures and its consequences in a system component or subcomponent parts. (Rochlin, La Porte and Roberts, 1987).

Another strategy that is used in high-reliability theory is that organizations that use this theory decentralize their authorities when making decisions. According to Rijpma, *"this is done to enable those closest to the problem at hand to solve problems as they emerge."* (Rijpma, 1997 p. 17).

However, the problem with this strategy is that people without decision-making authority and who have no knowledge of the problem at hand may not be able to effectively carry out the expectations of high-reliability theory. The proponents of HRT have negated the above-mentioned problem with the culture of reliability which inculcates clear operational goals and decision premises into members of the organization and allows for competence and autonomy when responding to complex interactions with high-reliability theory.

To achieve high reliability, enormous investments are made into expensive, elongated, painful trial and error learning processes and constant training is conducted before the standards are improved and the organization gets it right.

The strategies of the high-reliability theory were negated by Sagan, through his observation by stating the impossibility of anticipating every contingency of a failure or account for situations that are unexpected by the lower and higher levels existing within the organization's hierarchy. (Sagan, 1995). Thus, unexpected and unanticipated situations and interactions are bound to occur in complex systems.

Sagan and Perrow also negate the redundancy strategies in the high-reliability theory. According to them, redundancy in a complex system will increase the complexity of the system in more than one way. Besides, the system becomes opaque (Perrow, 2011; Sagan 1995) and components failure may become invisible and even become unnoticeable (Turner, 1978). Sometimes, the backup system serving as redundant may fail independently or simultaneously thereby adding to the complexity of the systems.

2.3. The Precautionary Principle

According to the 1992 Rio Declaration, the precautionary principle was defined as:

> *"To protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage,*

lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental damage" (UN 1992)

According to Renn, the principle is a highly contested strategy that relies on the principle that decisions should be carefully considered when there are (scientific) uncertainties. With precautionary principles, when uncertainties surround a decision, it is advisable not to execute the decision rather than encounter uncertainties toppled with negative consequences. (Renn, 2008).

Summary

The implication of the above discussed diverging views on literature review for this thesis is that LAWS is a high-risk technology that is complex in nature and tightly coupled. The lack of human intervention in LAWS makes the interaction between component parts and the subsystems highly interactive with little or no slack or buffer between two subsystems. What this means is that if a one part of the component system of LAWS is affected, other components parts are likely to be affected. Thus, normal accidents are inevitable yet unanticipated in LAWS. And due to the fact that the engagement and usage of LAWS have a ripple effect on human lives, this thesis views normal accident in LAWS from a larger lens because such normal accident can cause disasters and catastrophic events.

However, this thesis also views LAWS from HRT perspective which emphasis that even though normal accident is inevitable, it can be avoided with a decentralization of authority and redundancy strategy and argues whether safety measures for LAWS can be established through a HRT

Finally, precautionary principle is an important element that should be considered during deployment and engagement of LAWS. Thus, the precautionary principle among other principles that LAWS is expected to comply with will be addressed in this thesis.

CHAPTER 3

LETHAL AUTONOMOUS WEAPONS

3.1. UNDERSTANDING ARTIFICIAL INTELLIGENCE, THE UNDERLYING TECHNOLOGY OF LETHAL AUTONOMOUS WEAPONS

3.1.1 Artificial Intelligence

The existence of the Lethal Autonomous Weapon System is based on the application and use of A.I. AI is designed to replicate human behaviour intelligently while learning from experiences and improving the system's capacity to adapt. (Krishnan, 2021).

The United States, been at forefront of AI further defined it in the FY 2019 U.S. National Defense Authorization Act (NDAA) as:

> (1) Any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets.
>
> (2) An artificial system developed in computer software, physical hardware, or another context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action.
>
> (3) An artificial system designed to think or act like a human, including cognitive architectures and neural networks.
>
> (4) A set of techniques, including machine learning, that is designed to approximate a cognitive task.
>
> (5) An artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision-making, and acting. (§ 238 FY 2019 U.S. NDAA)

However, due to the different approaches that have been taken with regards to research in A.I, there seems not to exist a universal definition. Thus, due to imprecision, it is sometimes referred to as machine learning even though machine learning is a subset of AI.

According to Luxton D*., "the goal of AI is to build machines that are capable of performing tasks that we define as requiring intelligence such as for reasoning, learning, planning, problem-solving and perception."* (Luxton, 2016 p. 3). Therefore, the ability to build machines with intellectual capacity that is indistinguishable from humans and yet capable of further exceeding the capabilities of human intelligence is the goal of the use and application of A.I.

Research sponsored by NATO and led by Julian Lindley-French opined that:

> ….AI, deep learning, machine learning, computer vision, neuro-linguistic programming, virtual reality, and augmented reality are all part of the future battlespace. They are all underpinned by potential advances in quantum computing that will create a conflict environment in which the decision-action loop will compress dramatically from days and hours to minutes and seconds . . . or even less. This development will perhaps witness the most revolutionary changes in conflict since the advent of atomic weaponry and in military technology since the 1906 launch of HMS Dreadnought. The United States is moving sharply in this direction to compete with similar investments being made by Russia and China, which has itself committed to a spending plan on AI that far outstrips all the other players in this arena, including the United States. (Lindley- French 2017, 17).

Therefore, the lack of a universal definition of AI has had a ripple effect on the lack of a universal definition of Lethal Autonomous Weapons Systems (which will be discussed in subsequent chapters). This is because our inability to universally define A.I, the underlying technology of Lethal Autonomous Weapons Systems has affected how we define LAWS. Questions such as "Is LAWS AI?", "How do we define the lethality of LAWS?", "What makes LAWS autonomous?", "What is autonomy?", and a host of other questions have been raised for Lethal Autonomous Weapons System and yet, there seems to be no universal definition or responses to these questions.

3.1.2. Machine Learning

As mentioned earlier in this chapter, machine learning is a subset and core branch of AI. It allows computers to learn without having to be programmed with new information or data sets. (Samuel, 2000). It is recognized as the ability of a machine/software to enhance the performance of tasks through experience and data exposure.

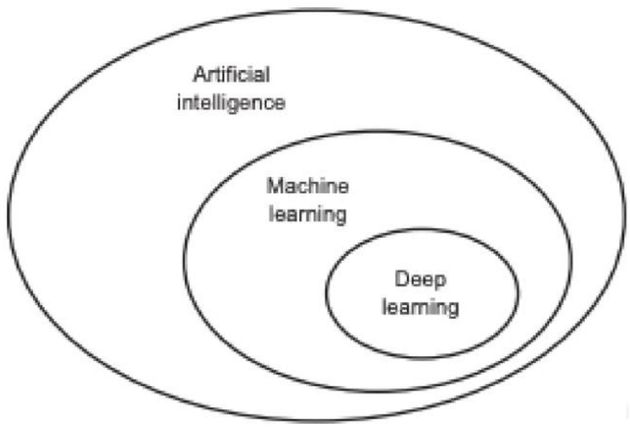

Figure 1: Correlation between AI, Machine Learning and Deep Learning (Chollet F., 2018)

Hence, the way a typical machine can improve its performance through machine learning will be to go through a learning process by being exposed to data and experience, learn from them and apply the knowledge acquired from the learning to make predictions and probabilities about emerging data. Sometimes these machine learning systems are supervised and trained with a pre-defined set of training examples/sets. Other times, the machine learning system is bestowed upon with the responsibility of discovering patterns and identifying relationships that exist in the data that it is exposed to, thus creating unsupervised learning.

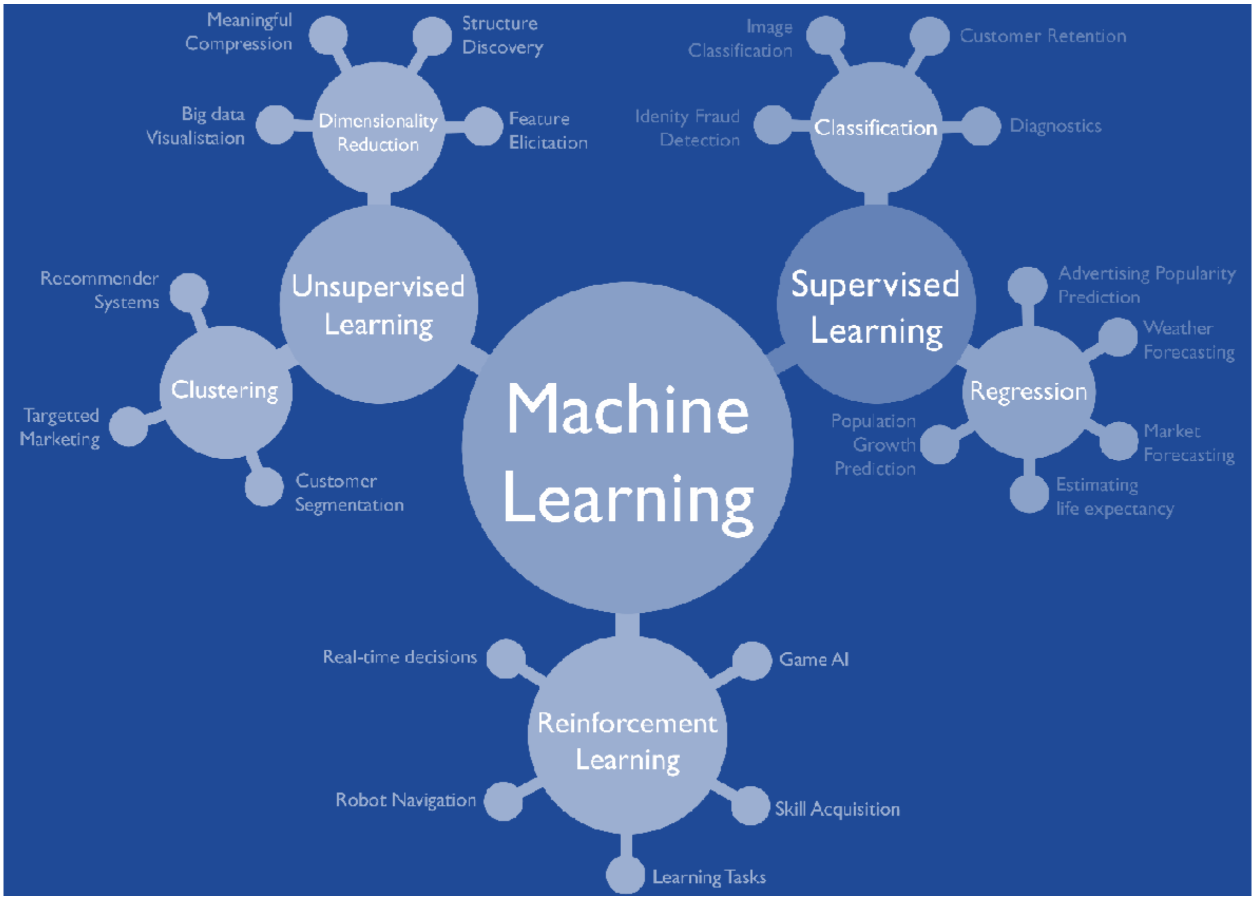

Figure 2: Machine Learning classification (Wahid, A., 2017).

Due to this complexity of machine learning, operators who have developed the system or running the system cannot comprehend the logic behind the systems. As such, it is challenging to produce evidence of formal proof explaining the behavioural pattern of a LAWS with a machine learning system or formally verify machine learning-based LAWS due to the requirements of formal proof expected of a software code-based system that performs critical functions that affect human lives. In addition, the stochastic nature of machine learning-based systems makes it difficult for developers, users, and operators to predict the LAWS' behaviour or explain the decision-making process.

3.1.3. Artificial Neural Networks

There is a type of machine learning that has been modeled after the human brain in the sense that they mimic the simulation, structuring, and functioning of the neuronal network in the brain.

In the brain, biological neurons make connections with each other through synapses, and they become strengthened over time through repeated use. It has been recognized as a mechanism used by the brain for acquiring, storing, and retrieving data and information efficiently and effectively and these connections *can "determine how neurons influence each other and how information flows in the form of a parallel processor*" (Alpaydin 2016, p. 86). Biological neural networks are simulated on computers by using neural network learning algorithms. These simulated modern neural networks use learning algorithms to make connections and where required, make adjustments to the connection weight between neurons. The way electrochemical impulses in the brain react to stimulants is the way artificial neurons react to data input received from sensors

According to Scharre,

> Neural networks do not perform rule-based calculations like most computers. Instead, they learn by exposure to large data sets. As a result, the internal structure of the network that generates output can be opaque to the designers—a "black box." Even more unsettling, for reasons that may not be entirely clear to AI researchers, the neural network sometimes can yield odd, counterintuitive results. (Scharre, 2016 p. 15).

Thus, LAWS operating on neural network are identified as black box systems with can arrive at engagement decisions without the developers, designers and/or the controllers understanding clearly how such decisions was reached. However, they are expected to be able to self-change their software or adapt their programming to produce better outputs and results just like a biological neurons network.

3.1.4. Deep Learning

Deep learning is a subset of machine learning but a larger aspect of Artificial Neural Networks. It uses the application of neural network algorithms to big data. Large sets of data that has been analyzed systemically and computationally to reveal patterns and trends (big data) are used to train the neural network to improve performance or solve a problem set such as recognizing differences or identify patterns and features that are most important for optimizing and refining distinctions and accurate predictions. The more data is fed to the computer systems for training, the better the output and results become.

As such, deep learning for LAWS is extremely useful when LAWS is deployed to environments that require LAWS to recognize differences or identifying patterns before performing its tasks. In fact, under the principle of distinction in IHL, it may be possible for deep learning LAWS to be able to make the necessary distinction for optimization and accurate predictions.

Deep learning allows for non-intuitive solutions of AI owing to the fact that during analysis and decision-making, AI is not limited by the cognitive bias of humans. This allows AI to think out of the box by operating above and beyond human cognitive limitations and produce better results and outputs. (Scharre, 2016). According to Kenneth Payne, Artificial Neural Networks *"are free from biological constraints and evolved heuristics, both of which serve to aid human decision-making amid pressures of time and uncertainty but can also produce systematic errors of judgment"* (Payne 2018, p. 171– 172).

Therefore, the combination of AI, machine learning, artificial neural networks, and deep learning is the underlying technology that forms Lethal Autonomous Weapons System. Without the afore-listed element, Lethal Autonomous Weapon Systems cannot exist, neither can they be developed or deployed for use.

3.1.5. The Turing Test

According to Luxton, *"any synopsis of AI would be remiss to not mention Alan Turing and the Turing Test."* (Luxton, 2016).

In 1950, Alan Turing, the brilliant British mathematician who was considered and recognized by others as the father of computer science (Beavers, 2013) designed and proposed a test, (which he called "the limitation game") that can measure and identify the differences and equivalence of behaviour between a machine and human. It is used as a means of judging the level of intelligence and the thinking capabilities of a machine. The test proposed is what is recognized now as the Turing Test (Turing, 1950). According to the test, a machine must be able to respond to queries/questions presented to it using the same structure that a reasonable human would use. If such a machine can fulfill that threshold, then the machine will be deduced to have the capability of intelligent thinking and may be assumed to have reached the level of AI. The Turing Test has become an inspiration in the AI history and community (Shieber, 2006).

Hence, before the deployment and use of Lethal Autonomous Weapon Systems, LAWS are assessed on the basis of a Turing test to judge the level of intelligence and capabilities. LAWS must pass the Turing test before it can be deployed.

3.1.6. The AI Black Box

A learning machine that doesn't follow a set of programmed or pre-defined rules but rather self-learn from data and adapts its programming to its environment is identified as BlackBox. With the black box, even though their conclusion seems accurately right or wrong, it is difficult if not impossible to understand why and how the system arrived at the decision or conclusion. As such, the ability to predict system failures in advance becomes extremely challenging if we do not understand the inner workings of the system.

AI researchers have identified that the functionality of a neural network is similar to a "black box" because the way computer codes and algorithms interact with big data makes it extremely difficult if not impossible for accountability and traceability. (Etzioni and Etzioni 2017, p. 35). With AI, we can see the output and result of the decision making but the process of decision-making which led to the arrival of the output/result is a mystery to human because it is difficult *and "one cannot easily follow the logic inside the hierarchical stack of artificial neurons that produced it"* (Payne 2018, p. 202) due to the complexity of interactions of the system and the feedback loops. (George 2003, p. 66).

According to a report in MIT Technology Review,

> …..you can't just look inside a deep neural network to see how it works. A network's reasoning is embedded in the behavior of thousands of simulated neurons, arranged into dozens or even hundreds of intricately interconnected layers. The neurons in the first layer each receive an input, like the intensity of a pixel in

> an image, and then perform a calculation before outputting a new signal. These outputs are fed, in a complex web, to the neurons in the next layer, and so on, until an overall output is produced. Plus, there is a process known as backpropagation that tweaks the calculations of individual neurons in a way that lets the network learn to produce the desired output. (Knight, 2017).

This leads to unpredictability and a lack of trust in the behaviour and decision-making of AI.

According to the scientific expert group called JASON studying the implication of AI for the U.S. Defense Department in 2017*, "the sheer magnitude, millions or billions of parameters (i.e., weights/biases/etc.), which are learned as part of the training of the net . . . makes it impossible to understand exactly how the network does what it does. Thus, the response of the network to all possible inputs is unknowable." (*Scharre, 2018 p. 189*)*

Hence, every time LAWS makes a decision to kill or not to kill, it may be difficult if not impossible to understand why and how LAWS arrived at such a decision. Therefore, if we cannot understand the process of arriving at a decision for LAWS, it might be challenging for us to predict system failure in advancing of it failing and causing accidents.

3.2. LETHAL AUTONOMOUS WEAPONS SYSTEM

3.2.1. What is Autonomy?

Autonomy has been viewed by Crootof to mean different things in different fields (Crootof, 2015). However, for the purpose of this thesis, Scharre's dimension of autonomy (Scharre, 2018 p. 34) will be analyzed. According to Scharre, some of the ways autonomy can be defined is in terms of the human-machine relationship and the sophistication of its decision-making process. Using the human-machine relationship, autonomy can be categorized into

i. Semi Autonomy (human in the loop)

A semi-autonomous system senses and observes the environment it is targeted to, and based on the circumstances and situations of the environment, the system relies on the pre-defined parameters and recommends the “appropriate” course of action to take. As such, the system cannot take a course of action without human approval. A simple and straightforward way to explain this is that the system senses and recommends, then it waits for a human to decide by either approving or rejecting the recommendation, then uses the human approval to act.

ii. Supervised autonomy (human on the loop)

Supervised autonomy is similar to full autonomy and semi-autonomy in the sense that the machine can sense, decide and act on its own without human intervention or command but a human user exists who sits on the loop to observe the machine’s behaviour and its result/output and where there is a need for a human user to intervene and stop the machine from executing its decision, such will be possible due to the existence of human supervision. (Scharre, 2018).

iii. Full Autonomy (Human out of the loop)

As the name indicates, these types of weapons have full autonomy and control over their operational decisions with the possibility to self-learn and adapt to new situations and information. (Davison, 2017). The system at this point “can sense, decide, and act entirely without human intervention” (Scharre, 2018). Thus, with humans out of the loop, the machine executes tasks without human user communication whatsoever. Also, their cognitive and decision-making process is unclear, untraceable, unpredictable, and less intelligible to humans because it is pretty difficult to understand how the system has performed the tasks due to its complexity. (Davison 2017 p.56).

Thus, a lethal autonomous weapons system can either be a semi-autonomous, supervised autonomous, or fully autonomous machine. Moreover, the type of autonomous can give insight and determine the level of risks such a machine could possibly face.

3.2.2 What is (Lethal) Autonomous Weapons?

A trip down the memory lane of history gives an evident view that the concept of autonomous weapons is not new. The path to autonomous weapons began sometime 150 years ago along the mid-nineteenth century and the concept has gradually evolved over a considerable period of time.

Although the “works of art” at that time cannot be considered fully autonomous, the idea behind these weapons can be seen to be one of autonomy. A practical example was the *“US-made Kettering “Bug” (a gyroscope-guided winged bomb) and the German FL-7 wire-guided motorboat, loaded with hundreds of pounds of explosives.”* (Lele, 2019 p. 53).

Lethal Autonomous Weapons Systems are recognized for disrupting the existing status quo of how warfare is fought and introducing reliability, precision, safety, and advancement to weapon systems. It has changed the perspective of countries, scientists, academia, developers, engineers, etc., on the way future battlefields should look like.

Currently, there are no settled or agreed definitions at the international level of Lethal Autonomous Weapons. (Borrie, 2016 p. 4). Definitions of Lethal Autonomous Weapons range from weapon systems to autonomous systems. One of the reasons is because it challenging to judge or identify what amounts to or can be regarded as “lethal”. However, according to Davison, autonomous weapons was defined *as ‘Any weapon system with autonomy in its critical functions—that is, a weapon system that can select (search for, detect, identify, track, or select) and attack (use force against, neutralize, damage or destroy) targets without human intervention.* (Davison, 2017, p. 5). This definition has also been accepted by the U.S. Department of Defense (DOD) and the United Nations.

According to the United Nations Special Rapporteur, Cristof Heyns warned *that “autonomous systems can function in an open environment, under unstructured and dynamic circumstances. As such their actions (like those of humans) may ultimately be unpredictable, especially in situations as chaotic as armed conflict, and even more so when they interact with other autonomous systems.”* (Heyns 2013, p. 8)

According to Scharre,

> A weapon system consists of a sensor to search for and detect enemy targets, a decision-making element that decides whether to engage the target and a munition (or other effectors, such as a laser) that engages the target.” (Scharre, 2018 p. 49). With autonomy, the weapon system can “finding, identifying, tracking, and prioritizing potential targets; timing when to fire; and maneuvering munitions to the target. (Scharre, 2018 p. 50).

Using the Advanced Medium Range Air to Air Missile (AMRAAM) as an example of a weapon system, critical analysis shows that it constitutes a radar, aircraft, pilot, and missile. With the help of the radar, the weapon system can scam, search and sense targets; for AMRAAM, a human decides whether to engage or not is required. However, in a system that has no human in the loop, the system decides whether to engage or not based on the parameters and its observations. Going back to AMRAAM, once a human decides to engage, the missile goes ahead to execute the engagement on the target.

The combination of the autonomy of the entire engagement loops and weapons system creates an autonomous weapon system. These weapons become autonomous because they can perform all these tasks, self-learn from the experiences, and re-program themselves to make better decisions when executing their tasks.

3.2.3. Types of Lethal Autonomous Weapon System

It is semiautonomous when a human intervention occurs and the human in the loop decides whether to engage the target or not. e.g., AMRAAM.

It is supervised autonomous weapons when the systems deployed and activated can perform their tasks without further human intervention, but humans are regarded to be "on the loop" because their function is to supervise the operation of autonomous weapons in real-time. Examples of supervised autonomous weapons are *"ship-based defenses, such as the U.S. Aegis combat system and Phalanx Close-In Weapon System (CIWS); land-based air and missile defense systems, such as the U.S. Patriot."* (Scharre P., 2018. p. 52). It is important to note that for supervised autonomous weapon systems, humans supervising are co-located physically with the systems and they monitor the operations in real-time. Even though there is no human intervention in terms of the decision-making and outcome of the system, humans are "on the loop" because they can intervene and physically disable the system where the need arises.

It is fully autonomous when the lethal weapon system can scan and search, make decisions to engage the target, goes ahead and engages the targets without any human intervention or supervision whatsoever. Lethal Autonomous weapon systems are not in wide use, but an example is the loitering munitions.

> *"Loitering munition is a complete "weapon system" all on its own. A human can launch a loitering munition into a "box" to search for enemy targets without knowledge of any specific targets beforehand. Some loitering munitions keep humans in the loop via a radio connection to approve targets before engagement, making them semiautonomous weapon systems. Some, however, are fully autonomous."* (Scharre P., 2018 p. 53)

The Israeli Harpy is another example of a fully autonomous weapon system. Harpy is an autonomous system that requires no human intervention.

Again, according to Scharre P.,

> The Harpy can stay aloft for over two and a half hours covering up to 500 kilometers of ground. This allows the Harpy to operate independently of a broader battle network that gives the human targeting information before launch. The human launching the Harpy decides to destroy any enemy radars within a general area in space and time, but the Harpy itself chooses the specific radar it destroys. (Scharre P., 2018 p. 55)

It is important to note that aside from using preprogrammed conditions and set parameters to automatically engage targets without any direct human control or intervention, Lethal Autonomous Weapon Systems have the capabilities to learn from experiences and their environments and use their new knowledge in form of datasets to improve their ability to function well with little or no human control. Thus, *"AWS would be able to go beyond their original programming and would reprogram themselves by optimizing desirable outputs."* (Krishnan, 2021).

Also, it is germane to correct the notion that the only thing that determines the level of a machine's intelligence is whether such a weapon is autonomous or not. Freedom of a system to make decisions without intervention is also determines an autonomous system and not intelligence only. Thus, a certain level of intelligence is required for a machine to freely make decisions to be called autonomous.

Hence, when a lethal autonomous weapon encounters a diverse and complex environment and there is a need to lock in a target in an organized crime syndicate, such a system will have to contend with irregularities that may occur, learning and self-learning, adapting to situations and environments and handling uncertainties based on strength of knowledge. Thus, a lethal autonomous weapon will have to predict the consequences of its action every time it engages a target in a circumstance that has not been encountered previously. This kind of processing technique, mechanism, and power that is given to autonomous weapon systems make the use of these weapons against humans especially organized criminals who look more or less like innocent bystanders a risky feat.

Lethal Autonomous Weapons having been recognized as a potential future weapon that can and would be used as a medium to engage targets independently without human intervention raises a lot of question and discussions about the risks involved in the use of such incredibly powerful weapon especially the consequences that may occur if system failure arises wherein LAWS engages a wrong/inappropriate target thereby leading to unexpected casualties and unintended consequences.

3.3. What is the security perspective of LAWS?

One of the most important questions circulating is that if LAWS becomes a standard in the international community, how will it affect stability? My view is that if the development, deployment and use of Lethal Autonomous Weapons leads to greater human control particularly in terms of how war is initiated and when the way wars are escalated, terminated and the reduction if not the elimination of the occurrence errors, miscalculations, failures, and accidents, there will be stabilization in the community. On the other hand, if the use of the Lethal Autonomous Weapons System leads to less human control, and there are accidents, failures, and unintended escalation, there is a viable possibility for instability in society.

It might sound counterintuitive to argue that the development, deployment, and use of autonomous weapons should lead to human control when the sole purpose and existence of LAWS is to reduce human control and delegate responsibilities and tasks to a machine, in this case, LAWS. However, increased human control in terms of the outcome is what is considered in this paper. Thus, when we allow a machine to perform a delegated task autonomously, human control over the result/outcome of the tasks can be increased. Using an example to explain this concept, when there is an imminent collision, the automobile collision avoidance system takes away control from the driver immediately. The control is relinquished to the system in order to achieve and attain the driver's desired result/outcome of avoiding a collision with another car, thereby increasing the driver's control.

However, even though the use of LAWS can increase human control and give stability, such stability can disappear and be replaced with instability due to the brittle nature of LAWS. The brittleness of LAWS can be seen as a major challenge especially when escalations cannot be controlled due to a lack of flexibility to adapt. A machine will do what it is set to do. Even though it has been argued that with AI, a machine can access and consider the information and data around it before making decisions, such flexibility will never match up to human's flexibility to adapt their decision to a specific event even though a different command has been issued in terms of the decision-making.

A practical example to support my argument is the disaster averted by Lieutenant Colonel Stanislav Petrov in 1983 when he ignored the information and notification alert from early-warning satellites implying that the United States had launched an unexpected attack. The Soviet automated missile alert system reported that the United States had launched five intercontinental ballistic missiles towards the Soviet Union.

Instead of the Lt. Colonel reporting the notification to the headquarters as required, he assumed and judged that a first strike consisting of only five missiles did not make any common and rational sense and was most likely an error or a glitch in the new system. Instead of reporting a U.S. attack, he reported that there seems to be malfunctioning in the system. It was subsequently found out that he had rightly assumed, and his judgments were right. False positives of "missile launches" from sunlight reflecting off of clouds (unanticipated interaction with the environment) were picked up by the Soviet Satellites as an incoming missile from the United States.

A similar incident to the above was the refusal of the Soviet Navy Captain Vasili Arkhipov to *"authorize the launch of a nuclear torpedo against the United States naval forces that were harassing a submarine under his command with signaling depth charges even though he was authorized to do so, and the submarine commander had ordered it"* (Hoffman, 1999).

These decisions prevented the occurrence of wars and gave humans increased control over the outcome of those decisions. A different scenario would have happened and could have escalated if these decision-making powers were given to a Lethal Autonomous System. LAWS would not have been able to read intuition, intents, and gut feelings. It would have set out the tasks it was ordered to and could have led to the beginning of wars. I would say that the inflexible and brittle nature of Lethal Autonomous Weapons could possibly take away human judgment which has been recognized as an important safety valve in crises and has led to the prevention of the occurrence of various wars. To know if LAWS can truly provide stability in society, it is important to address the issues of whether LAWS can comply with risk regulation and whether these risk regulations can provide stability. These debates will be discussed in chapter 5. However, before we discuss them, it is important we address the risks of Lethal Autonomous Weapons.

To summarize, LAWS cannot exist without A.I. However, there are different types of subsets of AI. They include machine learning, deep learning and neural networks. Thus, LAWS can either be developed and thoroughly trained through a machine learning, deep learning or neural network. The complexities involved in training LAWS either machine, deep learning or neural network makes it difficult for operators, developers and controllers to understand the logic and behavioural pattern of LAWS. After LAWS has been trained with one or more subsets, it must be assessed on the basis of a Turing test to judge the level of intelligence and capabilities. Thus, LAWS must pass the test before it can be deployed. Unfortunately, the Turing test does not explain the blackbox nature of LAWS, therefore making accountability and traceability of decision-making process extremely challenging for LAWS.

Even though the structural nature and the training modes of LAWS can give stability for safety, such stability can vanish and be replaced with instability due to the brittle nature of LAWS. Although, it has been argued that with AI, a machine can access and consider the information and data around it before making decisions, such flexibility will never match up to human's flexibility to adapt their decision to a specific event even though a different command has been issued in terms of the decision-making.

CHAPTER 4

THE RISK OF LETHAL AUTONOMOUS SYSTEM

4.1. Introduction to the practicality of Lethal Autonomous Weapon Systems

On March 23, 2003, the Tornado GR4A fighter jet over the north of Iraq was turned around by Kevin Main, British Lieutenant, and was headed towards Kuwait. At the back seat of the Tornado GR4A fighter jet, was the navigator, David Williams, flight Lieutenant. During the flight, it became unknown to Main and Williams that the identification friend or foe (IFF) signal that was supposed to be used to broadcast signal and notify other aircraft and ground radar that Tornado GR4A fighter jet was a friendly aircraft with no intention to fire was off. The IFF was not working and the reason for its failure to work is still mysterious up till today. It could have been from a possible power supply failure or the system itself failed. In fact, during maintenance, before the aircraft took off towards Kuwait, it was tested, and it should have functioned but for unknown reasons, it did not broadcast any signal.

As the Tornado GR4A fighter jet approached Ali Al Salem airbase, a radar signal was sent into the sky probing for Iraqi missiles and whether the approaching aircraft was a friendly or foe aircraft. The radar signal reflected on the fighter jet and bounced back to the base where the Patriot radar dish received the signal. Unfortunately, due to the fact that the IFF signal in the Tornado GR4A fighter jet wasn't on and couldn't broadcast a signal, the Patriot Computer failed to register the radar reflection from the fighter jet as a friendly aircraft. Also, because of the trajectory at which the aircraft was descending, the Patriot Computer tagged the radar signal (not the IFF signal) emanating from the fighter jet as an anti-radiation missile and the computer identified it as a radar hunting enemy missile. The humans in the loop were unaware that a friendly aircraft was descended for landing at the base.

The set parameter for the operation of the Patriot is to shoot down ballistic missiles. Anti-radiation missiles were never the Patriot's primary responsibility, but the Patriot had the authorization to engage and shoot down anti-radiation missiles if they appear to be homing in on the Patriot's radar.

The operators of the Patriot saw Tornado GR 4A fight jet as an anti-radiation missile headed and homing towards their radar and they had to weigh their decision within a split of seconds. In making that decision, the Tornado GR 4A fighter jet's IFF signal which should have shown on the Patriot's radar that a friendly aircraft was approaching was not broadcasting any signal.

Even if the IFF was working and it had broadcasted the signal, as it later turned out, the Patriots would not have seen or received the signal because the IFF codes were not loaded on Patriot's Computer. Thus, the Patriot Computer would not have identified the signal.

They had seconds to decide and in the mid of a split second, they took the shot and Tornado GR4A disappeared from the Patriot's scope. For those at the Ali Al Salem airbase and the operators of the Patriot, this was a hit, another success recorded as an enemy missile had been destroyed. To some others, Main and Williams landed in Kuwait at Ali Salem airbase. Unfortunately, the Tornado GR4A fighter jet never landed in Kuwait, the Patriot has shot down one of its own aircraft.

On April 2, another disaster struck. An inbound ballistic missile signal was picked up by a patriot operating north of Kuwait around Baghdad. There was no evidence or proof whatsoever that could suggest that the missile was misidentified. Since shooting down ballistic missiles was the primary responsibility of the Patriot, the launcher became operational, and the auto-fire system was engaged. The Patriot deployed Two PAC-3 missiles automatically.

However, what was unknown and oblivious to the Patriot and its operators was that there was no inbound ballistic missile. The signal picked by Patriot was caused by electromagnetic interference emanating from the Patriot's radar and another Patriot's radar nearby causing overlap and inference, thereby sending a false signal. But the operators and the Patriot itself could not have known.

The two PAC-3 missiles moved towards the spot where the incoming ballistic missile was supposed to be situated but the missile could not find anything. Unfortunately, a US Navy F/A 18c Hornet fighter was nearby the location where the radar was picked up. The PAC 3 missile from the Patriot locked onto the nearby aircraft. Even though the Hornet fighter jet sent an IFF signal which showed up on the Patriot's radar as a friendly aircraft and the pilot took some evasive action, PAC 3 still struck the aircraft within seconds thereby killing the Pilot, Lieutenant Nathan White.

The incidents above are practical examples of the risks of operating complex lethal autonomous weapon systems. The activities of the Patriots are a clear demonstration of the occurrence of Normal Accidents in a complex system and the complexity was a contributory factor to the operator's misconception and lack of comprehension of the Patriot's behaviour. Normal Accident will be further discussed in this chapter. However, we must address the effect of using lethal autonomous weapons.

4.2. Effects of using Lethal Autonomous Weapon Systems

The use of autonomous systems has caused significant controversy in discussions among policymakers, legal scholars, system developers, and risk professionals. However, when the conversation is moved up the notch to lethal autonomous weapons systems, an entirely different set of controversy, ignited disputes and table-turning evolves among the discussants.

While the use of fully autonomous weapons may not be common, there is proof of the existence of systems such as the US Aegis control system operating together with Phalanx Close-in Weapons System deployed to detect and neutralize possible missiles and aircraft. The South Korea armed robots, Samsung SGR-A are alleged to have been developed with an operating mode that allows you to select and engage targets without any human interaction or oversight. Moreover, it can identify human targets but cannot loiter freely neither can it create a distinction between friendly and hostile combatants. The Harpy Loitering Weapons of the Israeli is designed to identify and destroy radar emitters from the enemy. According to Scharre, "*it independently selects and attempts to destroy targets and only radars that meet the Harpy's programmed parameters will be engaged."* (Scharre, 2016 p. 7). A Naval Strike Missile with autonomous traits has been developed by Kongsberg Gruppen, the Norwegian defense company. The anti-ship missile uses algorithms to calculate and plan the best and accurate route to a target by identifying the best target and attacking the target at its weakest point.

The prospects and the risks of a full LAWS have raised multiple concerns on various fronts. However, amidst the multiple concerns, what are the potential security benefits of using a lethal autonomous weapons system?

It is my view that in an environment with a communication link that is broken down or poor, deploying Lethal Autonomous Weapons allows for the continuation of operation without a need for human communication to make decisions. In addition, compared to humans, the reaction time of autonomous weapons is quicker than humans, and this may be seen as an advantage in situations that quick and/or fast response is required for decision making. Also, LAWS are recognized to be reliable and with greater precision as compared to humans. Thus, it is safe to submit that despite multiple concerns, there are valuable benefits from the use of LAWS.

In furthering the argument of potential benefits of LAWS, it has been debated that the lack of human emotions with AI, the underlying technology for Lethal Autonomous Weapon can be seen as a benefit to LAWS because without emotions of bias, self-interest, loss, personal gain, etc., I tend to support that argument in the sense that AI devoid of emotions, bias and self-interest will be able to make objective choices and decisions during crises or when there is a need to make decisions.

When we view the possibilities and the limitlessness of LAWS especially when it is deployed to eliminate the high risk that comes with deploying humans and risking lives, it is safe to presume that it is a feasible replacement for combatants.

However, if Lethal Autonomous Weapons are deployed in situations that they have not been designed and tested for, where there is not enough data for the system to make its own decisions and reprogram itself and where there are changes to the set parameters in the environment, how then, will they be able to function effectively?

To answer this question, it is important to view the possibility of such systems failing or incapable of functioning effectively due to their inability to step outside the commands and instructions given to them and think out of the box in order to adapt to the circumstances they might find themselves. However, such possibility can be disputed in circumstances where AI has been designed and tested in such a way that they can step outside of commands and think out of this box, but this is a theoretical view that will hopefully see the light of the day sometime in the future.

LAWS are fundamentally different from any type of weapon because of their ability to self-act and self-determine thereby making them quite unpredictable. Due to Machine Learning and Artificial Neural Networks used in the development, LAWS can collate information from its environment, self-learn from experience or through probabilistic calculations where necessary and make independent decisions based on algorithms to arrive at a conclusion on how to act. However, the complexity of the weapon system makes it practically impossible for humans to predict the actions of the systems especially when the system is operating in a complex environment, and faced with malfunctioning, or acts in a peculiar way after gathering the necessary data and facts to make an expected decision.

4.3. The uncertainty and unpredictability of Lethal Autonomous Weapon Systems

Are LAWS unpredictable? LAWS have been identified to be unpredictable in terms of their decision outcome and result. Most schools of thoughts that support or oppose the use of LAWS have vehemently argued that the unpredictability of LAWS could be due to the opaque nature of AI which sometimes, makes it difficult for the system designers to understand how the system has arrived at a resulting action or decision because LAWS use artificial neural networks which have been designed to imitate the biological neural network to make decisions based on varying inputs and parameters.

Sometimes, feedback loops are created and become unknown to system operators due to the hidden interactions of the complex systems, and such cannot be detected. Thus, operators are faced with the challenges of understanding what these feedback loops mean or even realizing their existence in circumstances where their existence is outside the mental models of the workings of the system.

To further support the preposition on the unpredictability of LAWS, it is my view that the inevitable unpredictability that comes with using deep learning AI for LAWS is also another risk factor that has raised serious objections on the use and deployment of LAWS. A self-learning machine that uses numerous algorithms may sometimes produce inaccurate results due to the information bias that the system has been fed with. If the data the machine is trained with is limited to either a sector or a group, a weapon system with the deep learning capacity for image recognition, for example, might create a different perspective and understanding if it comes across an image different from the data it has been trained with. In addition, system designers cannot accurately predict the errors or malfunctions that may occur once LAWS is exposed to a real-life setting with an environment different from the lab where it had always been tested before exposure.

Thus, if a machine arrives at a decision through its unpredictable decision-making process, but the decision amounts to a loss of control of the LAWS, the autonomous weapon may continue to lose control without it knowing that it has lost control or made a mistake.

An important question that comes with unpredictability, is the uncertainty as to whether LAWS can make an accurate decision when faced with moral dilemmas. This question has been a subject of debate for human rights activists. Questions such as, does LAWS have the capability to go against orders and set parameters and choose not to target child soldiers even if the set parameters have no exception for child soldiers? How will LAWS make decisions for an organized criminal who has also been recognized as a civilian with a respectable position in society and with a loving family? or an offspring of a criminal who looks exactly like the target criminal? or an organized criminal who acts as an informant? These are the questions that developers have been unable to answer, and it is unclear how LAWS will make an accurate decision when faced with a moral dilemma. From my point of view, LAWS is incapable of making moral judgments but uses an algorithm based on probability to make judgments and not morality. Thus, a moral dilemma will be treated as a probable event with the best outcome. Unfortunately, using probability to make decisions in a moral dilemma situation might not yield the best result.

The element of uncertainty is another factor that should be considered when addressing the risks of deploying LAWS. The question that addresses the uncertainty of LAWS is whether a fully autonomous weapon can learn and observe the laws of armed conflict (LOAC). A practical example is when LAWS locks in a criminal as a target and the target surrenders, will LAWS recognize such target as nonbelligerent and therefore, disengages? The answer to the question of whether LAWS can learn and observe the laws of armed conflict (LOAC) will be discussed in detail in Chapter 5.

It is also important to note that the more complex the programming for LAWS is, the easier is it for a programming error to cause unanticipated events or unintended results.

4.4. System failure as Normal Accident with Lethal Autonomous Weapon Systems

According to Perrow's argument, Lethal Autonomous Weapon Systems are complex systems, and they are highly vulnerable to system failures under the Normal Accident theory due to unanticipated and unexpected interactions of the system in non-linear ways. (Perrow, 2011), Thus, when there is a rapid progression of system failure from one subsystem to another, and there is little or no slack in between these subsystems that can be used to either react to or absorb the failures, normal accident becomes inevitable.

To support Perrow's argument, it is my position that there are no fail-safe technology or error-free operations, especially with lethal autonomous weapons. Regardless of how careful the designers and operators have designed the system and operations; system failure is a common yet persistent risk that can occur in very complex technology systems. Due to the complexity and tight coupling of the lethal autonomous weapon systems or subsystems, system accident which is also called Normal Accidents are bound to occur.

If normal accidents are bound to occur, how easy is it to identify component failure that leads to a normal accident? It is my opinion that the complexity and the tight coupling of LAWS may make it absolutely unclear as to which of the components failed neither can the designers and operators anticipate. Even with careful designing, planning, and high reliability; unanticipated interactions that may lead to system failures are bound to occur once the systems are complex and tightly coupled. And events like the Third Mile Island, 1986 Chernobyl Nuclear reactor, the destruction of the NASA twin probes, and the 2010 Deepwater Horizon disaster tend to show that failure is bound to happen.

In elucidating the above theory, Charles Perrow (2011) accurately explains the challenges one might expect when anticipating and preventing the occurrence of an accident in a complex system. Many complex systems including the nuclear reactor at the three-mile island and the lethal autonomous weapons systems are tightly coupled. According to Scharre,

> *"Tight coupling is when interaction in one component of the system directly and immediately affects components elsewhere. There is very little "slack" in the system—little time or flexibility for humans to intervene and exercise judgment, bend or break rules, or alter the system's behavior. In the case of Three Mile Island, the sequence of failures that*

*caused the initial accident happened within a mere thirteen seconds. It is the combination of complexity and tight coupling that makes accidents an expected, if infrequent, occurrence in such systems. In loosely coupled complex systems, such as bureaucracies or other human organizations, there is sufficient slack for humans to adjust to unexpected situations and manage failures. In tightly coupled systems, however, failures can rapidly cascade from one subsystem to the next and minor problems can quickly lead to system breakdown.". (*Scharre 2018, p. 155)

Normal accidents are more likely to occur with LAWS especially where the system may have incomplete information or arrived at a wrong probabilistic decision which led to an accelerated and unanticipated pace of interactions between the system components and the targets.

Also, the inability of operators and developers to understand complex systems including how they behave makes it difficult if not impossible to prevent or manage normal accidents. The complexity of LAWS makes it practically difficult if not impossible to understand how the system arrived at its decisions and the processes or stages it took in arriving at that decision. This makes the system designers, and the users view LAWS as sometimes incomprehensible and opaque.

It is my submission that due to the high focus on accidents in a "high risks system", it might be safe to consider Normal Accident Theory as an excellent framework for LAWS and the possible implications when there is a failure in the system. Arguing for normal accident theory, the cause of a LAWS accident is likely to be found in the complexity and tight coupling of the system making it susceptible to failures. Individual failures in the subparts or subcomponents of the system might be trivial, however, when those subparts with trivial failure interact with one another, the failure becomes serious and such interaction of the multiple failures becomes recognized as an accident.

Using the Patriot fratricide as a practical example of how lethal autonomous weapons could behave, we can assume that the incidents that occurred on the Patriots were not mere occurrences but were failures that were impossible to anticipate and inevitable as a result of operating a highly lethal weapon system with high complexity and tight coupling thereby leading to Normal Accident.

The mere fact that the incident of F-18 and Tornado Fighter jet varied with different causes shows the extent of normal accidents occurring in complex and tightly coupled systems. However, there has been a question by the proponents of lethal autonomous weapons systems. The question is:

Is there any safety to the operation of complex and tightly coupled systems such as lethal autonomous weapon systems?

Under the Normal Accident Theory, accidents cannot be eliminated when it comes to complex systems that are tightly coupled. However, normal accident theory has been criticized and argued against by high-reliability theorists. The High-Reliability Theory is of the view that normal accidents can be avoided in certain circumstances if not eliminated.

4.5. High-Reliability Theory as a means of safety when using Lethal Autonomous Weapon Systems

In criticizing the normal accident theory, the proponents of High-Reliability Theory argue that "extremely safe operations are possible, even with extremely hazardous technologies, if appropriate organizational design and management techniques are followed" (Sagan, 1995 p. 13)

The US Navy's submarine community, the carrier flight decks, and the Federal Aviation Administration air traffic control systems have been recognized as examples of high-reliability organizations due to their ability to ensure a surprisingly low level of accidents even with the complexities of operating these systems. The US Navy established the Submarine Safety Programme (SUBSAFE) after the 1963 Thresher loss. (Scharre 2018, p. 164).

To support the high-reliability theory, Scharre argues that,

> *"Submarine components that are critical for safe operation are designated "SUBSAFE" and subject to rigorous inspection and testing throughout their design, fabrication, maintenance, and use. There is no silver bullet to SUBSAFE's high reliability. It is a continuous process of quality assurance and quality control applied across the entire submarine's life cycle. Upon installation and at every subsequent inspection or repair over the life of the ship, every SUBSAFE component is checked, double-checked, and checked again against technical specifications. If anything is amiss, it must be corrected or*

approved by an appropriate authority before the submarine can proceed with operations." (Scharre, 2018 p. 165-166).

For more than half a century, the US Navy has used the SUBSAFE programme without one submarine being lost. Ordinarily, this should be a myth under the Normal Accident Theory. However, SUBSAFE has become a practical example of how high-reliability theory can be used for safety when operating high-risk technology with complex systems and tight coupling.

The question that begs for attention and deserves a critical debate is could high-reliability theory be used to achieve safety when operating complex lethal autonomous weapons? My answer to the question is that using high-reliability theory seems not to be straightforward as it should be. High-Reliability Theory may be used to achieve safety for lethal autonomous weapons systems that have humans in the loop (either semi-autonomous or supervised autonomous) but it is unclear whether safety can be achieved for a fully autonomous weapon system that does not require human intervention.

To support the high-reliability theory and the principle of "human in the loop", the US Navy is of the opinion that, the track record and improvement of Aegis show that there is a possibility of using high-reliability theory for complex and tightly coupled systems. Testing systems are not sufficient for high-reliability theory to be effective, there need to be active participation of a human in the loop who will *"program the system's operational parameters, constantly monitor its modes of operation, supervise its actions in real-time, and maintain tight control over weapons release authority."* (Scharre 2018 p. 174). Moreover, the high-reliability operation is extremely costly and time-consuming and may prevent the international communities, governments/nations and, organizations from investing in such theory.

It is my view that high reliability may not be achievable because it requires that LAWS be tested frequently with various experiences and scenarios that may exist under real-life conditions. It is quite difficult to test a lethal weapon under a real-life scenario due to the unanticipated and unplanned events that come with a real-life scenario and the time and resources including money that will be invested in testing and testing over and over again.

However, Scharre, 2018 seems to disagree with my view. According to him, hazardous and complex systems have been managed safely by organizations such as the US Navy submarine and

Aegis because these organizations have embraced the High-Reliability Theory. He believed that if the high-reliability theory had been adopted earlier by the Patriot community, the incidents that happened to the Tornado and F/A 18c fighter jet could have been prevented either through cultural vigilance or testing which would have shown that electromagnetic interference between two radars can send an enemy signal. He argues that the high-reliability theory does support an accident-free situation but suggests the possibility of extremely low accident rates.

Even if we agree with Scharre's argument that high reliability supports accident-free situations and can eliminate normal risks in LAWS, can LAWS comply with risk regulation to bring about societal safety? The answer to this question shall be discussed in further detail in Chapter 5.

Assessing the risk of a complex system like lethal autonomous weapons can be extremely challenging. According to Feynman's observation in an appendix to the official report on the Challenger,

> *"It appears that there are enormous differences of opinion as to the probability of a failure with loss of vehicle and human life. The estimates range from roughly 1 in 100 to 1 in 100,000. The higher figures come from the working engineers and the very low figures from management. What are the causes and consequences of this lack of agreement? Since 1 part in 100,000 would imply that one could put a Shuttle up each day for 300 years expecting to lose only one, we could properly ask "What is the cause of management's fantastic faith in the machinery?"* (Feynman, 1986 para 1).

From Feynman's comments, we can draw some important conclusions surrounding the difficulties it entails when assessing risk in complex systems like lethal autonomous weapon systems. One of the conclusions I can draw is the difficulty in objectively quantifying the likelihood of the risk occurring and the risk itself. People will give varying opinions, views, and analyses of what they think and most of these may not be accurate judgment. Aggregating more data and testing these data can be useful in making accurate decisions and judgments on the likelihood of the risk but unfortunately, there is no straight-jacketed way of calculating or arriving at the likelihood of the occurrence of risk in a complex system like Lethal Autonomous Weapons.

4.6. Referring to Normal Accidents, would a weapon review of the Patriot System have made a difference in this circumstance?

It is unknown whether the Patriots' weapons in the above-mentioned circumstances were reviewed before their deployment for operation. However, it is important to debate whether a weapon review would have made a significant impact.

A weapon review can ensure that LAWS complies with the necessary risk regulation to reduce, and if possible, eliminate the risks that come with lethal autonomous weapons. With weapons review, we can create an environment where LAWS can be tested to ensure that LAWS can fare well when eventually deployed for functioning. The major challenge with creating such an ideal environment is the difficulty of getting an overview of the exact environment, the unforeseen circumstances surrounding it, and the challenges of knowing what will happen if LAWS is deployed.

However, relying on Additional Protocol I to the Geneva Convention, Article 36, there is an explicit obligation on States who are parties to the protocol to conduct weapon reviews. What then happens to states and private sectors not bound by the Protocol?

This provision creates a vacuum for States and private sectors who desire to use LAWS but are not bound by Article 36 of the Weapon Review. In relying on practices of other States before adopting Additional Protocol I, authors such as Bill Boothby have argued that there is "an applied obligation" to conducting weapons review (Boothby, 2016).

However, according to Boothby, evidence has shown that most States have no weapon review system and for those who have a weapon review system, such States neither comply with the legal obligations for weapon review. Sometime in 2006, a guide on weapons review was published by the ICRC stating that the United States, the U.K., Belgium, the Netherlands, Norway, Sweden, Australia, Canada, France, and Germany as the only members out of all the members of the Protocol who have developed a weapon review system. (ICRC, 2006). According to the Stockholm International Peace Research Institute in November 2015, "only a limited number of states (12 to 15) are known to have a weapon review mechanism in place." (Boulanin, 2015).

Even though Article 36 mandates States to conduct weapons review processes; the way, format, and manner in which such weapons review processes should be taken has not been articulated by

Article 36. Thus, conducting a weapon review process is seen by States as a national matter. This flexibility becomes an issue for standardization amongst State and private actors.

According to Article 36 of Additional Protocol I to the Geneva Conventions, states parties are required to ensure that in *the 'study, development, acquisition, or adoption of a new weapon'* that it is under an obligation to ensure that such party would not, *"in some or all circumstances, be prohibited by the Protocol or any other rule of international law"*. (Roff, 2015).

Article 36 has failed to be effective in addressing possible concerns about LAWS. Arguing against this would be a cause for astonishment owing to the fact that very few States have currently taken up the responsibilities to be actively engaged with weapons review processes; and where such a State is recognized to have actively carried out weapon review in practice, such actions have not *resulted in a prohibition of development, acquisition, nor adoption whatsoever.* (Grut, 2013 p. 21).

According to the ICRC, consideration technical performance of the weapon should be included as part of the review for new weapons. This is because technical performance is relevant in determining whether the usage and/or deployment of the reviewed weapon may cause indiscriminate effects, system failures, or accidents. (ICRC, 2006).

However, it is my view that no matter the amount of testing and verification of a system, unanticipated system failure is a likely occurrence, and a normal accident is inevitable.

CHAPTER 5

THE OBSTACLES LETHAL AUTONOMOUS WEAPONS SYSTEMS ENCOUNTER IN COMPLYING WITH RISK REGULATION

5.1. Introduction

The risk of regulatory challenges that truly autonomous weapons systems face arises from replacing humans' responsibilities with an autonomous system that is independent of human oversight or intervention. The question that should ordinarily follow this statement is, how is this a challenge?

The major problem is that once the human intervention is totally taken out of the loop, LAWS seems to be incapable of abiding by key principles of risk regulations such as the international humanitarian law and the laws of armed conflicts (LOAC). This is because regardless of the level and sophistication of LAWS, they are considered incapable of making the type of highly complex and contextual analyses that international humanitarian law requires.

The bedrock of risk regulation for Lethal Autonomous Weapon Systems, the international humanitarian law (IHL) has three core principles:

i. The principle of distinction means that only legitimate targets may be attacked. Thus, a clear distinction must exist between a civilian and an enemy combatant on the battlefield; and civilians cannot and must not be targeted deliberately. However, it has been acknowledged by IHL that sometimes, incidental killing may occur to civilians while targeting enemy combatants. This incidental killing is called "collateral damage."
ii. The principle of proportionality is to the effect that where collateral damage occurs, the civilian casualties should not be excessive or disproportionate to the necessity of attacking the target. (Scharre, 2018 p. 250)
iii. The principle of precaution means that during the processes of planning, execution, and deploying a lethal autonomous weapon, States must take precautionary measures at all times to minimize and if possible, eliminate and nullify the ripple effect of deploying lethal autonomous weapons on civilians and civilian objects.

What is the IHL's position on lethal autonomous weapons? According to the IHL principles in the Geneva Convention and the views of Scharre, principles such as distinction and proportionality have no effect on the decision-making process of a lethal autonomous weapon but on the effects of the decision-making process on the battlefield or place of conflict. (Scharre, 2018) It is also important to note that nothing in the laws of war prevents a machine such as LAWS from making decisions. However, for LAWS to be lawful and for proper risk regulation, LAWS must comply with the IHL principles of distinction, proportionality, and precaution among other rules.

However, Steve Goose, director of the Human Rights Watch's Arms Division and a leading figure in the Campaign to Stop Killer Robots, does not think that is possible. From his perspective, he sees autonomous weapons as *"highly likely to be used in ways that violate international humanitarian law"*. According to him, LAWS would be weapons *that "aren't able to distinguish combatants from civilians, not able to tell who's hors de combat, aren't able to tell who's surrendering, unable to do the proportionality assessment required under international humanitarian law for every individual attack, and that are unable to judge necessity in the way that today's human can."* The result, Goose said, would *be "lots of civilians dying."* (Scharre 2018, p. 251). The difficulty in meeting these criteria is dependent on the kind of technology, the environment that surrounds such technology, and the target.

5.2. Can lethal autonomous weapon systems comply with the principle of Distinction?

Two components of the principle of distinction (also referred to as discrimination) exist. According to Article 48 of the Additional Protocol I to the Geneva Convention, parties to armed conflicts are capable of distinguishing between civilians and enemy combatants. Also, they must be capable of distinguishing between civilian and military objects. (Henckaerts, 2009). To further buttress this, the International Court of Justice (ICJ) has opined that the principle of discrimination is a cardinal principle of international humanitarian law. (ICRC, 2006).

For proper compliance with this principle, the distinction between the civilian and military targets must be clearly and accurately distinguished by lethal autonomous weapons. What this connotes is that aside from being able to recognize the target, LAWS should be able to distinguish a target from a clutter of confusing objects similar to the target and surrounding the target's environment. According to Scharre, *"even for "cooperative" targets that emit signatures, such as radar, separating a signature from clutter can be challenging".* (Scharre, 2018 p. 251)

Thus, will LAWS be able to fulfill the criteria of distinguishing between civilians and enemy combatants during an attack?

In debating the question, on the surface, the principle of distinction seems quite simple and straightforward, like a black and white rule: a potential target is either a military or a civilian. Difficulties may arise however when it turns out that a target can be classified and take the profile of both military and civilian simultaneously, depending on the context from which the target is viewed. Besides, various electromagnetic signals ranging from cell tower signals to wifi routers, television, and radio broadcast signals are present everywhere in the modern urban society, and distinguishing such signals can be difficult. *It becomes extremely challenging to distinguish non-cooperative targets, such as tanks and submarines, that use decoys or try to blend into the background with camouflage*. (Scharre, 2018 p. 252) Thus, distinguishing people and their objectives, the far and most difficult task for a machine to embark on. In debating this issue for and against the ability of LAWS to distinguish, I will be addressing my discussion from the susceptibility of LAWS to "weak machine perception"?

According to the arguments of Petman, *"the kind of analysis that is generally required by the principle of distinction is a highly complex and highly contextual analysis that requires the uniqueness of the human mind to adept at".* (Petman, 2017 p. 28). Although deep neural networks are great at identifying and recognizing objects, they are still vulnerable to "image fooling" attacks and have not attained the highly contextual analysis that a human mind can perform. Thus, using LAWS for targeting in a highly contextual situation would be extremely dangerous (Scharre, 2018, p. 252) Also, LAWS can be programmed with rules such as "deploy missiles if fired upon," but in a complex and confusing situation, such a weapon may have been fired wrongly or unintentionally, LAWS been unable to identify intents may deploy missiles which could lead to unwanted fratricide.

Understanding human intent would require a machine with human-level intelligence and reasoning, at least within the narrow domain of warfare. (Scharre, 2018 p. 254) It is debatable whether LAWS can reach that pinnacle where it will be able to clearly distinguish civilians and its objects from combatants and its object with the same ability as humans. (O’Meara et al., 2011). Particularly, this is the situation in many asymmetric/lopsided conflicts and warfare that occur and are prevalent in today’s modern era. Thus, it can be challenging *to* distinguish a farmer digging a trench from a member of an armed group planting an improvised explosive device. (Petman, 2017); guerrillas wearing uniforms and/or civilian clothes interchangeably, or woodcutters and local farmers who include firearms as part of their farming tools to protect themselves from unwanted events or their property. In order to analyze and determine whether the above-mentioned are friends or enemies is dependent on their actions and even so, it is unclear whether LAWS can understand and contextualize their actions as opposed to a human. (Scharre, 2018 p. 252)

To support Steve Goose’s view, LAWS seems ill-equipped to make the required distinction between civilians and with sufficient or absolute clarity even though existing technology may have approached the capability level of distinguishing a human from a non-human object. Distinction is not a mere matter of LAWS sensors identifying the necessary signals, enemy weapons, or combatants, it is a matter of interpreting human behaviour by requiring a proper value judgment and a highly complex appraisal process.

For LAWS to properly make a distinction, LAWS will have to base its decision on understanding the intentions of humans and being aware of the situation in which it is targeting. Thus, situational awareness and human intentions are important factors for LAWS to comply with the principle of distinction. (Sharkey, 2010). Unfortunately, it is unclear and unknown whether LAWS is capable of identifying the above-mentioned two factors when found in a confusing environment. Addressing the expectations of LAWS in a confusing environment that requires situational awareness and human intention, some commentators have identified a specific criterion - ‘gut feeling’ that is common among humans and expressly included in the US army guidelines concerning ethical conduct on combat missions, which dictates that a soldier arriving at a decision whether the action he is about to take is morally “right” should be the final mental step that should be taken before deciding whether to shoot or not. (Geiss, 2015). Even advocates of LAWS during the Informal Expert Meeting on LAWS in 2014 acknowledged that such “gut feeling” deliberation may not be amenable and responsive to algorithmic programming. (Arkin, 2009). Thus, it is challenging and seems problematic to program rules of behaviors to combat situations with uncertainties. (Petman, 2017 p. 28).

To further support this proposition, in the nuclear weapons Advisory Opinion, the International Court of Justice postulated that *"states must……never use weapons that are incapable of distinguishing between civilian and military targets."* (ICJ, 1996 para 84). To support the position of the ICJ, Louise Doswald- Beck argues that *"it is obvious that a weapon, being an inanimate object, cannot itself make such a distinction, for this process requires thought." (*Doswald-Beck, 1997). According to Egeland, *"In the context of LAWS, however, attaining the technical capabilities demanded to be able to make these distinctions would be a necessary (but insufficient) condition for the lawful deployment of laws"* (Egeland, 2016). The ability to think through and identify human behaviour and intention is an important criterion to distinguishing between civilians and combatants.

5.2.1. Could the case be different if LAWS is used majorly to target other weapons?

It is my opinion that the case might be different, and LAWS may be able to make a proper distinction if they target other weapons systems only. To support my view, Phalanx is a practical example of an autonomous weapon designed only to shoot down missiles heading towards a ship, and it is unusual and to say the least extremely rare for Phalanx to mistakenly consider or view a civilian object on the ocean or approaching it for an incoming missile. Thus, LAWS programmed to fire at tanks has a significantly reduced possibility of firing or targeting civilian objects except LAWS finds it challenging to make a distinct decision between an active military tank at an exhibit or an abandoned tank in a civilian area. Petman (2017) has argued that targeting only weapon systems could limit the risk of mistaken targeting and enhance LAWS to comply with the principles of distinction. According to him,

> *"It is one thing to use an automatic anti-missile defense system in the middle of an ocean, it would be another thing entirely to target weapons systems in built-up civilian areas, or to target weapons that might not necessarily be military objects, such as civilian guns. In this respect, the risks of undermining the principle of distinction may be more serious in the case of LAWS that are mobile and increasingly capable of, say, directing their own flight paths, as compared to, for example, automatic sentry guns which could be placed only in well-marked areas where the presence of civilians is highly unlikely."* (Petman 2017 p.28).

On the other hand, without being biased, it can easily be argued that LAWS can perform a better job of discriminating targets and calculating the impacts of an engagement/task in real-time to ensure proportionality to the gained military advantage. Without emotion to impact LAW's decision-making processes, the ability of LAWS to make decisions becomes one of probability.

Based on probabilities, LAWS will not engage if there is a low probability of success or there is a high probability of success, but such success is linked to extremely high collateral damage. Also, in such situations where LAWS is uncertain about engaging a target, LAWS can be preprogrammed to ask questions and inquire from human further information to help make a better decision choice.

However, it is important to note that probabilities are not enough to discriminate and to disagree with this school of thoughts, it is necessary to bear in mind that a common trait that has engulfed many modern conflicts is that the employed tactics used in discriminating between civilians and combatants are a herculean task in practice. The fourth-generation warfare tactics are mostly used in recent wars. These so-called tactics include terrorism, guerrilla warfare, and immersion of combatants into the local population. (Hammes, 2004). The parties using these warfare tactics leech on the civilian's protections and display an element of surprise. Thus, such parties do not wear uniforms and they ensure that they blend themselves with the civilians without distinguishing themselves. This may make it very difficult for LAWS to distinguish them from civilians.

In the Stanford Law School's report on drones' strike, the selection of targets through the lens of a drone's camera was done with some tactics often based on some pre-defined/predetermined selection criteria such as activity, suspicious looks, location, or gender. It was therefore concluded that in targeting the presumed enemy, such targets were based on "signatures" that have been associated with some sort of terrorist activities (Egeland, 2016 p. 98):

> *Under the Obama administration, "signature" strikes that relied on "patterns of life" analysis were included in the drone program. According to the United States authorities, these strikes target "groups of men who bear certain signatures, or defining characteristics associated with terrorist activity, but whose identities aren't known"* (Cavallaro et al, 2012 p. 12)

In furthering the debates on LAWS, Noel Sharkey was of the opinion that for LAWS to be able to distinguish and differentiate between combatants and civilians, LAWS would be required to have not only attained the capacity to distinguish via sensory and vision processing but *"would have to overcome severe limitations in programming language and operationalizing common sense*" (Sharkey, 2012). The question of whether robots can successfully discriminate between combatants and civilians comes down to whether robots can think and function as human beings. The answer to that question is unknown yet.

5.3. Can LAWS comply with the principle of proportionality?

What the principle of proportionality seeks to achieve is to ensure that when LAWS is deployed for a specific attack, the collateral damage such deployment will cause to the civilians is proportional to the military advantage derived. Hence, the principle seeks to protect civilians. Where it is crystal clear that the collateral damage to the civilian is disproportional and excessive to the military advantage, such deployment should be prohibited to prevent civilian harm. (Roff, 2015).

For States to comply with the principle of proportionality, it is required that Parties to an armed conflict take all feasible precautions to avoid civilian harm and to ensure that targets are military objectives. (1125 U.N.T.S. 3).

In determining whether LAWS can meet the advantage requirement when making a decision as to proportionality and ensuring that the collateral damage to civilians is proportionate to the military advantage, one needs to ask if a reasonable human would have reached a similar decision/conclusion? To support this point of view of whether LAWS can meet the advantage requirement, the Trial Chamber of the International Criminal Tribunal for the former Yugoslavia was of the opinion that accountability for a disproportionate attack is dependent on the inherently 'reasonable human' standard. According to ICJ, *"whether a reasonably well-informed person in the circumstances of the actual perpetrator, in this case, LAWS, making reasonable use of the information available to him or her, could have expected excessive civilian casualties to result from the attack."* (Prosecutor v Galic, 2003*).*

When the military advantage concept of the principle of proportionality is viewed from the standpoint of "reasonable human standards", it opens opportunities for operational discretion and subjective assessment. Subjective assessment is an important element in adhering to the principle of proportionality. Hence, it is my submission that while humans are known to be capable of balancing complex interests on a case-by-case basis and make subjective assessments, the same cannot for the time be said of LAWS. Thus, the inability of LAWS to make decisions that are contextual and discretionary, two fundamental elements of proportionality, leads to concerns about whether LAWS are capable of complying with the proportionality principle.

To further support the argument on LAWS being unable to comply with the proportionality principle, there seems not to exist a clearly defined metric as to the requirements of the principle of proportionality. High contextuality is used when analysing proportionality. Besides, the requirement that an assessment of the level of civilian harm versus the value/benefits of the military advantage must exist and the need for considering other alternatives to gaining military advantage with lesser civilian harm makes the principle of proportionality and the ability to set a clearly defined metric a hard nut to crack.

However, based on the above-mentioned, the question therefore is, can LAWS be programmed to conduct such proportionality assessments? The principle of proportionality is quite complex and to adhere to this rule, LAWS would require to be programmed with a clear understanding of how to distinguish between civilian harm and military advantage gained and when civilian harm would be excessive to the benefit/value of a military advantage. To be able to make such a distinction and identify excessive civilian harm versus military advantage, LAWS would require special programming and understanding of military strategy, tactics, and operational issues. According to Egeland, *"LAWS would furthermore have to be able to comprehend continual changes in goals and objectives, internal changes to their relative importance and the anticipated military utility of achieving them."* (Egeland, 2016).

Applying and including the principle of proportionality to LAWS is "not an exact science" (Dinstein, 2016 p. 122), and programming the principle in the forms of codes into computer software will be even more challenging than expected (Asaro, 2013) especially if they have to be re-programmed for every mission or on a case-by-case basis.

However, arguing for the proponents of LAWS complying with proportionality, their view can be seen from the perspective of LAWS being able to predict the ripple effect of all potential decisions and the potentially resulting number of civilian casualties. (Wagner, 2011). Krishnan Armin in his book expressed that LAWS has the potential and ability to use force proportionately more than what a human soldier is expected to do. (Krishnan, 2016). With LAWS, there is an expectation of a more precise and quick calculation of the expected blast and the effect of other weapons which could cause collateral damage and whether such collateral damage is acceptable; *"while such calculations would be 'far too complex for the warfighter/human to make in real-time. LAWS could perform hundreds of these same calculations in real-time, increasing the lethality of the engagement while simultaneously reducing the probability of collateral damage"* (Guetlein 2005, 5).

To the best of my understanding, the above argument is flawed in the sense it may sound theoretically possible for LAWS to be programmed to comply with the principle of proportionality, practically, it seems practically impossible for LAWS to make qualitative and subjective decisions that require it to determine the excessiveness of the damage to civilian vis a viz the military advantage gained. (Sharkey, 2012). Besides, the principle of proportionality cannot be reduced or qualified to a clear-cut formula. (Solis, 2016). Thus, it is challenging for developers to write down such a proportionality principle into a piece of software coding. Even if developers could create a rule that is capable of being codified, it is unclear how LAWS will accurately determine the excessiveness and proportionality between civilians and the military.

Scharre (2018) faulted the argument by stating that it is challenging:

> ……..to have a machine decide about the proportionality of the attack. This would require the machine to scan the area around the target for civilians, estimate possible collateral damage, and then judge whether the attack should proceed. This would be very challenging to automate. Detecting individual people from a missile or aircraft is hard enough, but at least in principle could be accomplished with advanced sensors. How should those people be counted, however? …….Fighters who do not respect the rule of law will undoubtedly attempt to use civilians as human shields. How would an autonomous weapon determine whether or not

> people near a military target are civilians or combatants? Even if the weapon could make that determination satisfactorily, how should it weigh the military necessity of attacking a target against the expected civilian deaths? Doing so would require complex moral reasoning, including weighing different hypothetical courses of action and their likely effects on both the military campaign and civilians. Such a machine would require human-level moral reasoning, beyond today's AI. (Scharre, 2018, p. 255)

Determining proportionality is a judgment call and LAWS does not necessarily have the capability to make judgment calls without human intervention. With the current level of technological development, it is my submission that LAWS will never be able to satisfactorily make these types of determinations required, although the obstacles may presently seem conquerable.

5.4. Can the precautionary principle be used for LAWS to bring about societal safety?

In the 1907 Hague Convention (IX) concerning Bombardment by Naval Forces in Time of War, the principle of precautions during the attack was first set out. (The Hague, 1907). According to the principles, parties to an armed conflict are demanded and required to take certain precautions before initiating an attack.

This was also reiterated in Article 57 of Additional Protocol I to the Geneva Convention of 1977 that requires that:

> *'.....in the conduct of military operations, constant care shall be taken to spare the civilian population, civilians and civilian objects' do everything feasible to verify that objectives constitute military objectives; take all feasible precautions in choosing the means and methods of attack to avoid or minimize collateral damage to civilians and civilian objects and refrain from launching attacks expected to be in breach of the principle of proportionality*. (1125 UNTS 609).

Feasible precautions have been defined in the 1980 Convention on Conventional Weapons as *"those precautions which are practicable or practically possible taking into account all* *circumstances ruling at the time, including humanitarian and military considerations".* (1342 UNTS 137).

The principle of precaution extends not only to the commanders or controllers of LAWS but the manufacturers and programmers as well and it is inclusive of the planning phase of the deployment of LAWS. (Boothby, 2014).

The measures relating to this principle are extremely dependent on the context, prone to swift yet unpredictable change. Thus, the choice of weapons, selection criteria of targets, time, and method of attacking must be verified continuously to prevent any unexpected scenario. Where there is disproportionality in the impact, that is, if the civilian harm is excessive of the military advantage gained or the target is no longer a legitimate target due to surrendering, etc., engagement is expected to be canceled or suspended. Also, there is a need to take precautions by issuing prior warnings to the civilian population and community.

Owing to the fact that various unforeseen and unexpected circumstances and challenges are likely to occur during deployment or active combatant mission, some authors are of the perception and have argued that the precautionary principle has initiated the need for human intervention - always keeping humans 'on the loop' with supervisory control, thus giving human the ability to respond to any unexpected circumstances that may occur during deployment of LAWS. (Weizmann et al, 2014). Those of a different opinion believe that LAWS has the capability to process huge amounts of information and react to their environment more quickly than a human's response. These characteristics of LAWS question whether any human 'on the loop' in a combat situation has the capability to function quickly as LAWS or capable of early intervention and can prevent LAWS from violating an international humanitarian law principle. (Alston, 2011).

With the highly contextual nature and the need for a continuous reassessment as a requirement for the precautionary principle, many commentators are of the opinion and have concluded that LAWS cannot perform these required assessments continuously and independently without human intervention. (Herbach, 2012).

As another route to precaution, one may inquire whether the development and deployment of LAWS can be confined to exclude civilians from the outset and whether the decisions of LAWS relating to the principles of IHL should be executed by LAWS, or the execution be confirmed by a commander or controller that has the knowledge of the environment LAWS is deployed to and/or the authority to send LAWS into the field?

To answer the question, conflicts of today are transboundary yet non-international, combatants no longer define clearly the geographical battle lines. Military targets are known to be located in civilian vicinity with no clarity as to distinguishing combatants (military targets) from non-combatants (civilians). It is precisely due to these complex scenarios that LAWS cannot function independently without breaching the IHL principles and other risk regulations" (Geiss, 2015).

CHAPTER 6

DISCUSSIONS

Lethal Autonomous Weapons System poses a significant impact on the world's global security because of its limitlessness on the battlefield (Wan et al, 2018). Countries like United States, United Kingdom, Russia, China, France, Germany, and even Norway are currently investing heavily in the research and development of new applications using AI as an advantage over their counterparts.

Safety and security have been critical yet important issues facing the development and deployment of lethal autonomous weapon systems. A lethal autonomous weapon may comply with risk regulations and perform lawfully often. However, in the event that a system failure occurs or loss of control, such could become a disaster and catastrophic leading to civilian casualties and mass fratricides. Unfortunately, some of these system failures are rare but inevitable.

It seems challenging to mentally picture LAWS being able to carry out the required task for discrimination between combatants and non-combatants in attacks. For LAWS to adhere to the laws of war principle for risk regulation, it is required that the execution of qualitative judgments by LAWS is based on data and information which may be challenging to quantify. As such, the rule of proportionality is similarly predicated on qualitative interpretation, and there is an obligation to assess whether the humanitarian, material and environmental consequences of an attack seem justifiable viz a viz the military advantage gained. Where there is a change in the dynamics of the equation during the course of an attack, it is difficult and unknown to see how LAWS can take such occurrence into account.

In terms of weapons review, States and the international community have expressed their concerns and questioned the processes for legal review. Their concerns revolve around whether *"all States had all the necessary and required technical expertise to implement the process of weapons review effectively". (*Meier, 2016). Other States have expressed their concerns on the application of different national standards of weapons review due to lack of standardization, and if such different standardization can be sufficient to ensure that LAWS is consistent and compliant with International Humanitarian Laws. (U.N., 2015). In response to the concerns of States, the United

States Delegation proposed that the Convention on Certain Conventional Weapons should step in and create a document detailing a comprehensive process of weapon reviews. Such a document becomes a standard used for common approach and understanding and will be applied by any State with the intention of developing, acquiring, or deploying LAWS.

Ordinarily, a responsible military in a country will not launch to the battlefield a weapon that has been identified to have a high risk of failure or that seems to be burdened with unknown consequences. However, developing and deploying lethal autonomous weapons goes against the ordinary status quo and amounts to accepting the consequences of inevitable system failures of a high-risk technology such as LAWS. Thus, how should States and the international community address the development and deployment of Lethal Autonomous Weapon Systems? What strategies should they engage in?

From the above arguments, one thing is clear, the risk regulation principles that bring about societal safety and security cannot be adhered to by a fully autonomous lethal weapon system without human intervention. In light of this, there has been enormous pressure from the international civil society and other anti-LAWS advocate on various States and governments to *"ban killer robots."* However, there are practical challenges to banning LAWS. First and foremost, the difficulties in reaching a consensus on the definition of a lethal autonomous weapons system make it difficult to determine what is LAWS and what isn't under banning circumstances. In addition, the government is not the only actor when it comes to the research and development of LAWS. Private companies are the major actors in the research and development of LAWS and it might be difficult for a democratic government to place a ban on LAWS.

Thus, what can be done to address the risks and uncertainties of the development and use/deployment of lethal autonomous weapon systems?

In addressing the risk and uncertainties of developing and deploying LAWS, an option could be to rely on the precautionary principle. Precautions can be taken by identifying LAWS based on the risks of accidents occurring and other possible risks and expressing that the risks are way too enormous for LAWS to be developed and deployed, thereby prohibiting the use of LAWS and totally eliminating the risks and uncertainties that come with developing and deploying LAWS. According to the UNIDIR resources, *"...this might be a legally binding prohibition, or be achieved by means of an international moratorium until the international community agrees on clear* *standards for determining how accidents would be prevented, and how agreed standards would be verified." (*Borrie, 2016*).*

Another alternative that may be highly useful in verifying and evaluating LAWS is to create a framework that encompasses effective technical standards and necessary procedures that should be complied with by States who are interested in the development and deployment of LAWS. In addition, protocols that will assist end users with the required information and training, guide risk assessments, and point out other fail-safe measures should be created for LAWS.

However, the likelihood and possibility of developing a framework for LAWS for weapons review are quite unclear. Also, while it is important to explore and address the possibilities of creating fail-safes, creating such fail-safes for LAWS can be likened to a herculean task that seems insurmountable due to the reasons highlighted in this thesis.

Regardless, the risks LAWS *"are diverse, and they demand an approach that adequately identifies and controls the risk arising from behavioural uncertainty and generates information that can be used to better evaluate, manage and prevent the risk of unlawful behaviour"* (Bhuta et al, 2016 p. 299).

Alternatively, LAWS can be limited to a semi-autonomous weapons system. This type of weapon system will require human intervention, that is, a human in/on the loop to authorize LAWS engagement and act as a natural fail-safe in the event of possible failures and accidents.

Where LAWS engagements go contrary to the principles of international humanitarian law, the human controller who is recognized as the human in the loop can alter or halt the operations of the system before further damage is wrecked. Moreover, with a human in/on the loop, LAWS may comply with the risk regulations requiring human intervention and judgment.

With a fully autonomous weapon system, the potential damage that may be caused and its effect before a human intervenes could be far greater than expected. In extreme events where the failure mode or contradiction to risk regulation is replicated in other LAWS of the same type, society could likely face a catastrophic disaster where a large number of developed or deployed LAWS fail simultaneously.

LAWS can be advantageous in situations where precision, speed, and reliability are required among other things, but it is also important to note that LAWS are brittle. They lack the flexibility of humans in adapting to novel situations. Their lack of flexibility gives us an insight into what a fully autonomous weapon system would have decided on if faced the same event Lt. Colonel Stanislav Petrov faced on September 26, 1983? My guess is as good as yours, it would have carried out the tasks that it was programmed to do without the option of human judgment or gut feeling and could have caused the beginning of the end of wars especially during the period of high Cold War tensions.

CHAPTER 7

CONCLUSIONS

It is important that government and civil society understand LAWS and their implications through a normal accident theory framework. Once the system is understood from a normal accident theory, then it is important to also understand LAWS from a high-reliability theory. Once both views have been reconciled. This will give some preliminary insights, guidance, and steps that can be taken, and countries and the international community need to make effort to ensure that LAWS developed and deployed have undergone a proper weapons review, comply with risk regulation principles of discrimination, proportionality, and precaution under international humanitarian law.

To answer the question whether LAWS can be viewed from the perspective of Normal Accident. It is clear from the arguments canvassed that LAWS are complex and tightly coupled systems, and they are highly vulnerable to system failures caused by unanticipated and unexpected interactions of the system in non-linear ways. These system failures called normal accidents are bound to occur and inevitable.

Individual failures in the subparts or subcomponents of the system might be trivial, however, when those subparts with trivial failure interact with one another, the failure becomes serious and such interaction of the multiple failures becomes recognized as an accident. Even with careful designing, planning, and high reliability; unanticipated interactions that may lead to system failures are bound to occur.

To respond to the question whether high reliability theory can be used as a means of safety for LAWS, the argument that extremely safe operations are possible, even with extremely hazardous technologies, if appropriate organizational design and management techniques are followed will not hold water for LAWS because achieving safety operations for a complex and tightly coupled system is not that straightforward and simple. Moreover, it is capital intensive and time consuming, and it requires frequent testing under real life scenarios to operate a HRO. Hence, private sectors whose focal point is to make profits and contribute to innovation may not attempt high reliability theory while developing, deploying and engaging LAWS. Hence, HRT cannot be used to achieve the desired safety.

A weapon review under Article 36 of the Additional Protocol I to the Geneva Conventions may bring society safety in the deployment and use of LAWS. However, most States and private actors have no weapon review system. Besides, there is no standardization for the way, format and manner a weapon review process should be. This may lead to different standards and measures used by States and private actors to review weapons, and this may ultimately threaten safety in the international community. Until a standardization of weapons review which includes technical performance is established, weapons review cannot bring about safety.

In addressing the security perspective of LAWS, **e**ven though the structural nature and the training modes of LAWS can give stability for safety, such stability can disappear and be replaced with instability due to the brittle nature of LAWS.

LAWS are fundamentally different from any type of weapon because of their ability to self-act and self-determine thereby making them quite unpredictable. Due to machine learning and Artificial Neural Networks used in the development, LAWS can collate information from its environment, self-learn from experience or through probabilistic calculations where necessary and make independent decisions based on algorithms to arrive at a conclusion on how to act. However, the complexity of the weapon system makes it practically impossible for humans to predict the actions of the systems especially when the system is operating in a complex environment, and faced with malfunctioning, or acts in a peculiar way after gathering the necessary data and facts to make an expected decision. This makes it difficult for LAWS to be recognized as a machine that can bring about safety in the society.

For LAWS to bring about society safety in the international community, LAWS must comply with the IHL and LOAC principles. Currently, there are obstacles for LAWS in complying with risk regulation. First and foremost, LAWS seems to be incapable of abiding by key principles of risk regulations such as the international humanitarian law and the laws of armed conflicts (LOAC) once the human intervention is out of the loop.

Furthermore, LAWS cannot comply with the principle of distinction because the kind of analysis required by the principles of distinction are highly complex and contextual analysis, situational awareness and human intentions that is only unique to humans. LAWS do not have a human-level intelligence to make complex and contextual analysis that will allow it to successfully distinguish between a civilian and combatants in complex environments. Besides, situational awareness and human intentions cannot be codified as an algorithmic programming.

In addition, LAWS cannot comply with the principles of proportionality in the sense that balancing interest on a case-by-case basis, making contextual and discretionary decisions are some of the fundamental elements of proportionality. Unfortunately, LAWS are incapable of executing these fundamental elements of proportionality. Even if we can codify these elements into LAWS, they are quite complex and such principles cannot be reduced to a clear-cut algorithmic formula or written down into a piece of software coding by developers.

Also, with the highly contextual nature and the need for a continuous reassessment as a requirement for the precautionary principle, it is challenging for LAWS to perform these required assessments continuously and independently without human intervention.

Although "normal accident" is to be expected with LAWS, this should not prevent us from creating norms and developing standards to ensure LAWS safety. While normal accident theorists are pessimists and the majority of them advocate a ban on LAWS, where norms and standards have been developed to ensure LAWS safety, none of them would advocate avoiding or neglecting safety and procedure. Thus, normal accident theory should remind us of all of the extent of "limits to redundancy, safety devices and organizations and the continuous existence of limitations to human's capacity to correctly predict problems/challenges and possible failures.

Furthermore, private industries should be involved in discussions relating to risk regulation of LAWS and compliance to IHL principles. According to Peter Singer, he is of the opinion that *"...... not a single organization, research lab, or company working on robotics today is formally linked up with the [International Committee of the Red Cross] or has in place… reviews… necessary for new weapons." (Singer, 2009).*

Efforts are being made by the International Committee for Robot Arms Control to involve the private sectors more in discussions because joint efforts between the public and private sectors are required when discussions are around risk regulations and international humanitarian principles. Having a joint discourse from the private sector, humanitarians, governments and other international communities are extremely important for the successful development, deployment, and compliance with risk regulations and the principles of international humanitarian law. Thus, as LAWS development arises, continuous diplomacy and discussions that cut across all sectors must exist.

States have been known to have a solid and longstanding history of cooperation among one another especially in terms of innovation and development of common norms. Thus, there should be a continued international, mutual dialogue and cooperation between States and private actors for the unified development of common norms and standards in the deployment and use of LAWS. This is seen as a necessary tool useful in managing potential strategic risks and possible dangerous outcomes that LAWS may face from the development, deployment, and use of it.

Conclusively, the international community needs to be prepared and face the reality that the development and deployment of LAWS do not fit into the existing framework of risk regulations governing weapons. It is extremely difficult and quite unfortunate that nothing can be done about that for now. However, the Normal Accident Theory has proved to be a useful and reliable framework that can be explored in understanding the challenges to LAWS, especially the fact that systems do fail in unexpected and unanticipated ways. With the international community having a better understanding of these issues, we can all begin to take the necessary steps to address the plights that LAWS will likely present.

REFERENCES

BOOKS

Anderson, K., & Waxman, M. C. (2017). Debating Autonomous Weapon Systems, their Ethics, and their Regulation under international law in Brownsword, R., Scotford, E., & Yeung, K. (Eds.). (2017*). The Oxford handbook of law, regulation, and technology*. Oxford, Oxford University Press. 1097

Alpaydin, E. (2016). *Machine learning: the new AI*. Massachusetts, MA. MIT press

Arkin, R. (2009). *Governing lethal behavior in autonomous robots*. CRC Press.

Beavers, A. (2013). Alan Turing: Mathematical mechanist. In S. B. Cooper, & J. van Leeuwen (Eds.), *Alan Turing: His work and impact* (pp. 481-485). Waltham, MA: Elsevier

Bekey, G. A. (2005). *Autonomous robots: from biological inspiration to implementation and control*. MIT press.

Bhuta, N., Beck, S., Geiβ, R., Liu, H. Y., & Kreβ, C. (Eds.). (2016). *Autonomous weapons systems: law, ethics, policy*. Cambridge, NY. Cambridge University Press.

Boothby, W. H. (2016). *Weapons and the law of armed conflict*. Oxford, Oxford University Press.

Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. New York. WW Norton & Company.

Chollet, F. (2018). *Deep learning with Python* (Vol. 361). New York: Manning.

Dennett, D. (1985). Can machines think? In M. Shafto (Ed.), *How we know* (pp. 121-145). San Francisco, CA: Harper & Row

Dinstein, Y. (2016). *The conduct of hostilities under the law of international armed conflict.* Cambridge, NY. Cambridge University Press.

Georges, T. M. (2003). *Digital Soul: Intelligent Machines and Human Values*. Boulder, CO: Westview Press.

Hammes, T. X. (2006). *The sling and the stone: on war in the 21st century*. St. Paul, MN, Zenith Press.

Haykin, S., & Network, N. (2004*). Neural networks: A comprehensive foundation*. New Jersey, Prentice Hall.

Krishnan, A. (2016). *Killer robots: legality and ethicality of autonomous weapons*. U.K. Routledge.

Krishnan, A. (2020). Enforced Transparency: A Solution to Autonomous Weapons as Potentially Uncontrollable Weapons Similar to Bioweapons. Lethal Autonomous Weapons In Galliott, J., MacIntosh, D., & Ohlin, J. D. (Eds.). (2020). *Lethal Autonomous Weapons: Re-Examining the*

Law and Ethics of Robotic Warfare. Oxford, Oxford University Press. Available at https://oxford.universitypressscholarship.com/view/10.1093/oso/9780197546048.001.0001/oso-9780197546048-chapter-15. Accessed on the 10th of April 2021.

Lin P., Bekey G. & Abney K. (2009). 'Robots in War: Issues of Risk and Ethics' in R. Capurro & M. Nagenborg (eds) *Ethics and Robotics* 49– 67, at 57–58. Heidelberg, Verlag

Luxton, D. D. (2016). *An introduction to artificial intelligence in behavioral and mental health care. In Artificial intelligence in behavioral and mental health care* (pp. 1-26). Academic Press.

Mitchell, T. M. (1997*). Machine learning*. Burr Ridge, IL: McGraw Hill.

Newell, S., & Simon, H. A. (1956). *The logic theory machine: A complex information processing system*. Santa Monica, CA: The Rand Corporation.

O'Meara, R. M., Lin, P., Abney, K., & Bekey, G. (2011). *Contemporary governance architecture regarding robotics technologies: an assessment. Robot ethics: The ethical and social implications of robotics.* p. 159-68. Cambridge, MA, MIT Press

Payne K. (2018). *Strategy, Evolution, and War: From Apes to Artificial Intelligence*. Washington, DC: Georgetown University Press

Perrow, C. (2011). *Normal accidents: Living with high-risk technologies*-Updated edition. Princeton, N.J., Princeton university press.

Renn, O, (2008), *Risk Governance: Coping with Uncertainty in a Complex World*, London, Sterling, VA, Earthscan.

Roberts K.H., (1993), "Introduction" in Roberts K.H. (Ed.), *New Challenges to Understanding Organizations*, pp. 1-10. New York, Macmillan.

Sagan S.D. (1995), *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*, Princeton, NJ. Princeton University Press

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. New York. WW Norton & Company

Shanahan, M. (2015). *The technological singularity.* Massachusetts, MIT press.

Singer, P. W. (2009). Wired for war: The robotics revolution and conflict in the 21st century. London, Penguin.

Snook, S. A. (1996). *Practical drift: the friendly fire shootdown over northern Iraq*. Harvard University.

Snook, S. A., & Friendly, F. (2000). *The accidental shoot down of US black hawks over northern Iraq.*

Solis, G. D. (2016). *The law of armed conflict: international humanitarian law in war.* Cambridge, New York. Cambridge University Press.

Turner, B.A. (1978), *Man-made Disaster*. London. Wykeham

Weizmann, N., & Costas Trascasas, M. (2014). *Autonomous weapon systems under international law*. Geneva Academy of International Humanitarian Law and Human Rights (Geneva Academy). Available at www.geneva-academy.ch/our-projects/publications

JOURNALS

Amikura, H. (2003). Scott A. Snook, Friendly Fire: The Accidental Shootdown of US Black Hawks over Northern Iraq. *Organizational Science Tokyo*, *37*(2), 75-76.

Alston, P. (2011). Lethal robotic technologies: the implications for human rights and international humanitarian law. *Journal of Law Inf. & Sci.*, *21*, 35.

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross*, 94(886), 687-709.

Cavallaro, J., Knuckey, S., & Sonnenberg, S. (2012). Living under drones: Death, injury, and trauma to civilians from US drone practices in Pakistan. *International Human Rights and Conflict Resolution Clinic.*

Crootof, R. (2015). "The Killer Robots Are Here: Legal and Policy Implications." *Cardozo Law Review*, 36, 1837.

Doswald-Beck, L. (1997). International humanitarian law and the Advisory Opinion of the International Court of Justice on the legality of the threat or use of nuclear weapons. *International Review of the Red Cross* (1961-1997), 37(316), 35-55.

Egeland, K. (2016). Lethal autonomous weapon systems under international humanitarian law. *Nordic Journal of International Law*, 85(2), 89-118.

Etzioni, Amitai and Oren Etzioni. (2017). "Should Artificial Intelligence Be Regulated*?" Issues in Science and Technology* 33 (4): pp. 32– 36.

Geiss, R. (2015). The international-law dimension of autonomous weapons systems. *Friedrich-Ebert-Stiftung International Policy Analysis.*

Grut, C. (2013). The challenge of autonomous lethal robotics to International Humanitarian Law. *Journal of Conflict and Security Law*, *18*(1), 5-23.

Henckaerts, J. M. (2009). The ICRC Study on Customary International Humanitarian Law: Characteristics, Conclusions and Practical Relevance. *Slovenian L. Rev.*, 6, 225.

Herbach, J. D. (2012). Into the Caves of Steel: Precaution, cognition and robotic weapon systems under the International Law of Armed Conflict. *Amsterdam LF*, *4*, 3.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *IEEE Computer Society Press*, 29(3), 31-44.

La Porte, T.R. and Consolini P.M. (1991), "Working in Practice but Not in Theory: Theoretical Challenges of "High Reliability Organizations", *Journal of Public Administration Research and Theory*, Volume 1, Number 1, Winter, pp 19-47

Lele, A. (2019). Debating Lethal Autonomous Weapon Systems ". *Journal of Defence Studies*, 13(1), 51-70. See https://idsa.in/jds/jds-13-1-2019-debating-lethal-autonomous-weapon

Meier, M. W. (2016). Lethal Autonomous Weapons Systems (LAWS): Conducting a Comprehensive Weapons Review. *Temp. Int'l & Comp. LJ*, 30, 119

Morkevicius, V. (2014). Tin men: ethics, cybernetics and the importance of soul. *Journal of military ethics*, *13*(1), 3-19.

Pedron, S. M., & da Cruz, J. D. A. (2020). The Future of Wars: Artificial Intelligence (AI) and Lethal Autonomous Weapon Systems (LAWS). *International Journal of Security Studies*, 2(1), 5

Petman, J. M. (2017). Autonomous Weapons Systems and International Humanitarian Law:'Out of the Loop'?. *The Eric Castren Institute of International Law and Human Rights*.

Rijpma, J. A. (1997). Complexity, tight–coupling and reliability: Connecting normal accidents theory and high reliability theory. *Journal of contingencies and crisis management*, 5(1), 15-23.

Rochlin, G.I. La Porte, T.R. and Roberts K.H. (1987), The Self-Designing High Reliability Organization: Aircraft Carrier Flight Operations at Sea" *Naval War College Review*, Volume 40, Number 4, Autumn, pp. 76-90

Roff, H. M. (2015). Lethal autonomous weapons and jus ad bellum proportionality. *Case W. Res. J. Int'l L.*, 47, 37.

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 44(1.2.), 206-226.

Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and Machines*, 10, 463-518.

Schmitt, M. (2013). Autonomous weapon systems and international humanitarian law: a reply to the critics. *Harvard National Security Journal*, 4, 1-37.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, 3(3), 417457. Available from: http://dx.doi.org/10.1017/S0140525X00005756.

Sharkey, N. (2008). Cassandra or false prophet of doom: AI robots and war. *IEEE intelligent Systems*, *23*(4), 14-17.

Sharkey, N. (2010). Saying NO to lethal autonomous targeting. *Journal of military ethics*, 9(4), 369-383.

Sharkey, N. E. (2012). The evitability of autonomous robot warfare. *Int'l Rev. Red Cross* 94, 787.

Singer, P. W. (2012). Wired for war: The robotics revolution and 21st century conflict. *Stosunki Międzynarodowe International Relations*, 46(2).

Turing, A. M. (1950). Computing machinery and intelligence. mind lix (236): "460. bona fide field of study. he has cochaired the aaai fall 2005 symposium on machine. *IEEE Intelligent Systems*, 2.

Wagner, M. (2011). Taking humans out of the loop: implications for international humanitarian law. *Journal of Law Inf. & Sci.*, 21, 155.

Wan, J., Yang, J., Wang, Z., & Hua, Q. (2018). Artificial intelligence for cloud-assisted smart factory. *IEEE Access*, 6, 55419-55430

REPORTS AND PROCEEDINGS

Boulanin, V. (2015). Implementing Article 36 weapon reviews in the light of increasing autonomy in weapon systems. Stockholm International Peace Research Institute [Stockholms internationella fredsforskningsinstitut] (SIPRI).

Davison, N. (2017). A legal perspective: Autonomous weapon systems under international humanitarian law. UNODA Occasional Papers No. 30.

Feynman, R. P. (1986). Appendix F: Personal observations on the reliability of the shuttle. Report of the Presidential Commission on the Space Shuttle Challenger Accident, 2, F1-F5.

Guetlein, M. A. (2005). Lethal autonomous weapons--ethical and doctrinal implications. NAVAL WAR COLL NEWPORT RI JOINT MILITARY OPERATIONS DEPT.

Heyns, C. (2013) "Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions." United Nations General Assembly: A/HRC/23/47. April 9, 2013

International Committee of the Red Cross, A. (2006). Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977, January 2006. IRRC, 88(864), 933.

International Committee of the Red Cross (2011): "International Humanitarian Law and the Challenges of Contemporary Armed Conflicts," Report of the 31st International Conference of the Red Cross and Red Crescent. Geneva, p. 39, https://perma.cc/ML3F-RUZB

Kellenberger, J. (2011). International humanitarian law and new weapon technologies. 34th Round Table on Current Issues of International Humanitarian Law.

Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, I.C.J. Reports 1996, p. 226, International Court of Justice (ICJ), 8 July 1996, available at: https://www.refworld.org/cases,ICJ,4b2913d62.html [accessed 11 June 2021]

Lindley- French, J. (2017). "One Alliance: The Future Tasks of the Adapted Alliance." GLOBSEC NATO Adaptation Initiative. November. https://www.globsec.org/wp-content/uploads/2017/11/GNAI- Final-Report-Nov-2017.pdf

Shieber, S. M. (2006, July). Does the Turing test demonstrate intelligence or not?. In Proceedings of the National Conference on Artificial Intelligence (Vol. 21, No. 2, p. 1539). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

United Nations Office at Geneva, Convention on Certain Conventional Weapons, Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS) ¶ 55, CCW/MSP/2015/3 (June 2, 2015), http://www.unog.ch/80256EDD006B8954/(httpAssets)/587A415BEF5CA08BC1257EE0005808FE/$file/CCW+MSP+2015-03+E.pdf [hereinafter 2015 Chairs Report

LEGAL INSTRUMENTS AND CASE LAW

International Committee of the Red Cross (ICRC), *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I)*, 8 June 1977, 1125, UNTS 3. Available at: https://www.refworld.org/docid/3ae6b36b4.html [accessed 11 June 2021]

International Committee of the Red Cross (ICRC), *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II)*, 8 June 1977, 1125, UNTS 609. Available at: https://www.refworld.org/docid/3ae6b37f40.html [accessed 11 June 2021]

International Conferences (The Hague), Hague Convention (V) Respecting the Rights and Duties of Neutral Powers and Persons in Case of War on Land, 18 October 1907. Available at: https://www.refworld.org/docid/3ddca4e14.html [accessed 11 June 2021]

Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II), June 8, 1977, 1125 U.N.T.S. 609

Second Protocol to The Hague Convention of 1954 for the Protection of Cultural Property in the Event of Armed Conflict, The Hague, 26 March 1999, entry into force 9 March 2004, UNESCO Doc. HC/1999/7, Art. 7.

United Nations, (1980) Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects (and Protocols) (As Amended on 21 December 2001), 10 October 1980, 1342 UNTS 137. Available at: https://www.refworld.org/docid/3ae6b3ac4.html [Accessed 7 June 2021]

§ 238 FY 2019 U.S. National Defense Authorization Act (NDAA)

1992 Rio Declaration on Environment and Development: UN Doc. A/CONF.151/26 (vol. I), 31 ILM 874

Prosecutor v Galic (Judgment) ICTY-98-29 (5 December 2003), 43 International Legal Materials 794 at 808

ARTICLES, WEBSITES AND OTHERS

Borrie, J. (2016). Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies. UNIDIR Resources, 5.

Carvin, S. (2017). Normal Autonomous Accidents: What happens when killer robots fail?. Available at SSRN 3161446 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3161446

Headquarters, U. S. Air Force, United States Air Force Unmanned Aircraft Systems Flight Plan, 2009-2047 (Washington, DC: Headquarters US Air Force, 18 May 2009). Available at www.fas.org/irp/program/collect/uas_2009.pdf at 379

Hoffman, D. (1999). I had a funny feeling in my gut. Washington Post, 10. See http://www.washingtonpost.com/wp-srv/inatl/longterm/coldwar/shatter021099b.htm.

Horowitz, M. C., & Scharre, P. (2015). Meaningful Human Control In Weapon Systems. See https://www.jstor.org/stable/pdf/resrep06179.pdf

Hubbard, C. (1997). A Critique of the Advisory Opinion of the International Court of Justice on the Legality of the Threat or Use of Nuclear Weapons, 8 July 1996: The Nuclear Weapons Case. Available https://ro.ecu.edu.au/cgi/viewcontent.cgi?referer=https://scholar.google.no/&httpsredir=1&article=1684&context=theses_hons

Knight W. (2017). "The Dark Secret at the Heart of AI." MIT Technology Review. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Marijan B., (2016) "On Killer Robots and Human Control," Ploughshares Monitor 37, no. 2. See http://ploughshares.ca/pl_publications/on-killer-robots-and-human-control (Accessed April 20, 2021).

Pryor, P., & Capra, M. (2012). Foundation Science. HaSPA (Health and Safety Professionals Alliance), The core body of knowledge for generalist OHS professionals. Tullamarine, VIC: Safety Institute of Australia.

Remarks by Thomas Nash, Director, Article 36, Informal Expert Meeting on Lethal Autonomous Weapon Systems, Convention on Certain Convention of Weapons, Geneva (2014) at 51.

Scharre, P. (2016). Autonomous weapons and operational risk. Available at https://www.files.ethz.ch/isn/196288/CNAS_Autonomous-weapons-operational-risk.pdf

Statement, Michael W. Meier, Head of Delegation, U.S. Delegation Closing Statement and the Way Ahead to the Convention on Certain Conventional Weapons (CCW) Informal Meeting of Experts on Lethal Autonomous Weapons (LAWS) (Apr. 17, 2015), https://geneva.usmission.gov/2015/05/08/ccw-laws-meeting-u-s-closing-statement-and-the-way-ahead/

The bans by the Campaign to Stop Killer Robots, http://www.stopkillerrobots.org/the-solution/ Human Rights Watch: https://www.hrw.org/topic/arms/killer-robots, and the International Committee for Robot Arms Control http://icrac.net/call/

Wahid, A., (2017), Big data and machine learning for Businesses, retrieved from: https://www.slideshare.net/awahid/big-data-and-machine-learning-for-businesses