MILITARY ROBOTS: ETHICS OF LETHAL AUTONOMOUS WEAPON SYSTEMS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SALİH GÜLMEZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF ARTS
IN
THE DEPARTMENT OF PHILOSOPHY

OCTOBER 2023

Approval of the thesis:

MILITARY ROBOTS: ETHICS OF LETHAL AUTONOMOUS WEAPON SYSTEMS

submitted by **SALİH GÜLMEZ** in partial fulfillment of the requirements for the degree of **Master of Arts in Philosophy, the Graduate School of Social Sciences of Middle East Technical University** by,

Prof. Dr. Sadettin KİRAZCI
Dean
Graduate School of Social Sciences

Assoc. Prof. Dr. Aret KARADEMİR
Head of Department
Department of Philosophy

Assoc. Prof. Dr. Barış PARKAN
Supervisor
Department of Philosophy

Examining Committee Members:

Assoc. Prof. Dr. Aziz F. ZAMBAK (Head of the Examining Committee)
Middle East Technical University
Department of Philosophy

Assoc. Prof. Dr. Barış PARKAN (Supervisor)
Middle East Technical University
Department of Philosophy

Assoc. Prof. Dr. Sibel KİBAR KAVUŞ
Kastamonu University
Department of Philosophy

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Salih GÜLMEZ

Signature:

ABSTRACT

MILITARY ROBOTS: ETHICS OF LETHAL AUTONOMOUS WEAPON SYSTEMS

GÜLMEZ, Salih
M.A., The Department of Philosophy
Supervisor: Assoc. Prof. Dr. Barış PARKAN

October 2023, 85 pages

In this thesis, the ethical impacts of Lethal Autonomous Weapon Systems (LAWS) have been investigated. The main focus of the thesis is the question of whether LAWS lead to a responsibility gap. The responsibility gap argument posits that no one bears responsibility for the actions of LAWS, resulting in a gap in responsibility assignments. However, I introduce the concept of vicarious responsibility, demonstrating that designers of LAWS can be held morally responsible for their design due to their moral entanglement. The central argument of the thesis posits that it is possible to attribute moral responsibility, albeit in a vicarious sense, to the designers of LAWS, thereby bridging the responsibility gap.

Keywords: Moral Responsibility, LAWS, Ethics of AI

ÖZ

ASKERİ ROBOTLAR: ÖLÜMCÜL OTONOM SİLAH SİSTEMLERİNİN ETİĞİ

GÜLMEZ, Salih
Yüksek Lisans, Felsefe Bölümü
Tez Yöneticisi: Doç. Dr. Barış PARKAN

Ekim 2023, 85 sayfa

Bu tezde, Ölümcül Otonom Silah Sistemlerinin (OSS) etik etkileri araştırılmıştır. Tezin ana odak noktası, OSS'lerin bir sorumluluk boşluğuna yol açıp açmadığı sorusudur. Sorumluluk boşluğu argümanı, OSS'lerin eylemleri için hiç kimsenin sorumlu tutulamayacağını ve bu nedenle sorumluluk atamalarında bir boşluk oluşacağını iddia eder. Ancak, bu çalışmada, tasarımcılar ve tasarımları arasında özel bir ahlaki bağlantı olduğunu söyleyen vekaleten sorumluluk kavramı ortaya atılmış ve tasarımcıların OSS'lerin eylemleri için sorumlu tutulabileceği öne sürülmüştür. Sonuç olarak, tezin ana argümanı sorumluluk boşluğunun ortadan kalkacağını iddia eder.

Anahtar Kelimeler: Ahlaki Sorumluluk, OSS, YZ Etiği

To my BG

ACKNOWLEDGMENTS

I am deeply indebted to my thesis advisor, Assoc. Prof. Dr. Barış Parkan, whose guidance and patience have been invaluable. I would also like to thank my thesis committee members, Assoc. Prof. Dr. Aziz F. Zambak and Assoc. Prof. Dr. Sibel Kibar Kavuş, for their thoughtful insights and constructive feedback.

I want to thank my parents, Sevilay and Mustafa, and my sister, Cansu, for their psychological and financial support. To countless friends, both old and new, thank you for providing much-needed moments of joy during this time. I would like to especially thank Zeynep for standing by me through both the ups and downs of life. I would also like to thank Ümmü, my library comrade, without whom the countless hours spent in the library would have been unbearable. Thank you for bringing fun and snacks. To Yogi, Régine, and Henri, for putting a smile on my face even in times of despair.

Lastly, I would like to thank Bastien for his encouragement. Without him, this thesis could have never been completed.

TABLE OF CONTENTS

APPENDICES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AP	Additional Protocols to the Geneva Conventions
CCW	United Nations Convention on Certain Conventional Weapons
DARPA	Defense Advanced Research Projects Agency
DoD	The United States Department of Defense
IAI	Israel Aerospace Industries
ICRC	International Committee of the Red Cross
IHL	International Humanitarian Law
LAWS	Lethal Autonomous Weapon Systems
ML	Machine Learning
NATO	The North Atlantic Treaty Organization
NGO	Non-governmental Organization
OODA	Observe-Orient-Decide-Act
UK	The United Kingdom
UN	The United Nations
US	The United States of America

CHAPTER 1

INTRODUCTION

The term "Artificial Intelligence" was coined in 1956 during a renowned workshop at Dartmouth College. Half a century later, in 2006, the fiftieth-anniversary event of the historic workshop was also held at Dartmouth College. Both workshops, the original and its commemoration, shared a common feature. Apart from the apparent commonality between the workshops, i.e., artificial intelligence (AI), another less prominent feature of both workshops is that both were sponsored by the Defense Advanced Research Projects Agency (DARPA) (Bringsjord & Govindarajulu, 2022).

DARPA is a governmental agency in the United States with a primary focus on research and development of defense technologies. On its official website, DARPA outlines its mission as "to make pivotal investments in breakthrough technologies for national security" (DARPA, n.d.). Although DARPA is not a military agency per se, its sponsorship of both workshops on artificial intelligence suggests a clear interest in the potential of artificial intelligence for military applications.

Lethal Autonomous Weapon Systems (LAWS) have attracted significant attention in ethical discussions about military uses of AI. However, one should distinguish LAWS from remotely operated drones, which are already extensively used in modern warfare. Unmanned aerial vehicles like drones are not considered lethal autonomous weapon systems, although they may share similar technological characteristics. The function of the drones is "to navigate, but not select and engage targets, *autonomously*" (Lele, 2017, p. 59, emphasis added). Drones rely on human operators to undertake lethal actions. Remotely operated systems, such as drones, are called *human-in-the-loop* systems because human operators play an essential role, particularly in lethal decisions. In contrast, LAWS removes the need for "human judgment in the initiation of lethal force" (Asaro, 2012, p.693). Thus, LAWS differ from unmanned aerial vehicles in their ability to operate without direct human control behind their actions. Drones have already increased the physical distance between humans and military operations on the battlefield by enabling human operators to control these systems remotely. LAWS are expected to increase this distance further. Once deployed, the actions of LAWS will no longer depend on a human operator's direct control or supervision. Unlike existing systems, which can perform some functions through remote control, artificial intelligence allows LAWS to have the capability to adapt to their environment, learn from it, and detect targets without human intervention.

On the other hand, these capabilities come with drawbacks. AI, particularly machine learning (ML), differs from traditional programming techniques. In traditional programming, the input is processed using a fixed algorithm defined by the programmer. Thus, it is known by the programmer which input results in which output. In contrast, systems equipped with machine learning are provided with a vast amount of data, and the system generates its algorithm from the training data (Alpaydın, 2016). Because the system generates the algorithm, the programmer may not know the exact procedure by which it processes input to output. This lack of transparency means that predicting the system's output in unforeseen situations becomes difficult due to the inherent complexity of machine learning systems.

This feature of new systems equipped with machine learning, i.e., unpredictable behavior, is conceived as a fundamental problem when LAWS are deployed in warfare. Thus, the activists, NGOs, and academics call for an international ban on LAWS. They argue that the deployment of LAWS is both unethical and unlawful. Three significant problems surrounding the deployment of LAWS have attracted attention: the principle of discrimination, the principle of proportionality, and the gap in responsibility (Asaro, 2012; Bartneck et al., 2021). Thus, in this thesis, I will analyze the ethical problems pertaining to the use of LAWS. After the preliminary considerations concerning the compliance of LAWS to the principles of *jus in bello* and IHL, special attention will be given to the problem of responsibility within the context of LAWS. Moral responsibility is an important factor because if the machine will make life-and-death decisions, then it becomes important who bears responsibility for the consequences of these decisions. To discuss these issues, in what follows, I will present the structure of the thesis. In Chapter 2, I analyze the various definitions and frameworks used to comprehend the concept of autonomy in LAWS. This analysis includes the often-cited definitions of LAWS provided by the US Department of Defense (DoD) and the International Committee of the Red Cross (ICRC). In addition to these definitions, the chapter explores the widely-used loop framework, which categorizes autonomous weapon systems based on the level of human involvement, such as human-in-the-loop, human-on-the-loop, and human-out-of-the-loop. The thesis specifically focuses on human-out-of-the-loop systems, where the decision-making procedure is fully delegated to the system's algorithm. Furthermore, the chapter delves into the programming foundations of autonomy in machines. For this purpose, a comparison between rule-based programming and machine learning has been given. Ultimately, Chapter 2 aims to provide a clear and comprehensive understanding of LAWS, which serves as a beneficial foundation for understanding the ethical issues in subsequent chapters.

Chapter 3 of the thesis focuses on the ethical concerns related to LAWS, particularly their compliance with International Humanitarian Law (IHL). One should distinguish between two ways of understanding "ethics of LAWS": one that concerns the ideas under the field of machine ethics, and the other concerns ethical discussions on the use of LAWS. The former is the study of implementing ethical decision-making into robots by creating artificial moral agency. The latter refers to the ideas on the ethical use of LAWS by investigating whether LAWS would comply with the ethics and laws of war. Although I will briefly discuss the former, as it will appear, the present thesis mainly pertains to the latter sense of ethics of LAWS, i.e., the investigation of the ethical questions surrounding the use of LAWS in warfare.

Keeping the above distinction in mind, Chapter 3 of the thesis proceeds as follows: I will first briefly describe just war theory. Particular attention is given to the ethical issues surrounding LAWS' compliance with two pivotal principles of IHL and just war theory: the principle of proportionality and the principle of distinction.

Chapter 4 focuses on the problem of moral responsibility assignment in unethical conduct of LAWS. The chapter starts with a general overview of moral responsibility. Then, I analyze the ethical problem known as the "responsibility gap" that emerges in situations where there is an ethically significant conduct of LAWS; however, no one is morally responsible for this conduct. The gap in responsibility occurs because no individual in the design, development, and deployment stages of LAWS has direct control over the actions of LAWS. After analyzing the responsibility gap argument, I propose that there is a sense of moral responsibility that allows the possibility of holding designers responsible. This sense of moral responsibility is known as vicarious responsibility, where one agent is responsible for the actions of another because of the morally relevant connection the two have. Vicarious responsibility is notoriously unclear because it aims to justify a moral connection that the traditional understandings of moral responsibility could not easily explain. To overcome the uncertainty and obscurity about this notion of moral responsibility, I use a modified version of the formal definition given by Glavanicova & Pascucci (2022). The formal analysis, I believe, helps to mitigate the obscurity inherent in vicarious responsibility.

Consequently, in this thesis, I argue that the responsibility gap problem can be overcome. Vicarious responsibility allows an analysis of how we can hold the designers morally responsible for the moral harm caused by the system they create. After arguing for this view, I also elaborate on why the argument that designers should be held responsible could be of use to other proposed solutions for the responsibility gap. These alternatives propose that responsibility rests with the collective as a group agent or should be distributed among individuals participating in the design, development, and deployment phases of LAWS. In conclusion, I argue that designers' vicarious responsibility also serves as a moral justification for the aforementioned alternative solutions: collective and distributed responsibility.

CHAPTER 2

WHAT IS LETHAL AUTONOMOUS WEAPON SYSTEMS (LAWS)?

There have been several attempts to highlight the importance of a clear and agreed-upon definition of LAWS (e.g., Crootof, 2015; Taddeo & Blanchard, 2022b). However, despite these efforts, the debate surrounding the definition of LAWS continues to occupy a significant portion of the literature. It should be noted that while there are similarities in the definitions proposed, there are also differences that complicate the discourse on LAWS. This lack of consensus is an essential factor that hinders the ability to draw conclusions regarding the ethical and legal implications of the LAWS.

As a matter of fact, there is as yet no uniformity on how to name these weapon systems. Future of Life Institute, for example, calls these systems “slaughterbots” (n.d.). Various terms such as “lethal autonomous weapon systems,” “autonomous weapon systems,” “lethal autonomous robots,” “killer robots,” and “fully autonomous weapon systems” are also employed to describe weapon systems with different levels of autonomy (Vilmer, 2015). This diversity in nomenclature further adds to the confusion surrounding LAWS.

In its 2012 Directive (updated in 2023), the US Department of Defense (DoD) describes an autonomous weapon system as “a weapon system that, once activated, can select and engage targets without further intervention by an operator” (2023, p. 21). According to DoD, independence from the intervention by a human operator plays a crucial role in defining these systems. Although the definition of DoD is one of the most prominent definitions in the literature, it might also be confusing when discussing about these systems. The independence from human operator is a crucial part of lethal autonomous weapon systems. However, it is insufficient to describe these systems as merely capable of performing independently of human operators. DoD's definition, at first glance, seems precise and straightforward, but it falls short in defining what the acts of selecting and engaging targets entail. In other words, it is not clear what LAWS are performing when they are performing without human operators.

Another commonly cited definition of LAWS was given by the International Committee of the Red Cross (ICRC). ICRC defines LAWS as "any weapon systems with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e., use force against, neutralize, damage or destroy) targets without human intervention" (ICRC, 2016, p. 1). These two definitions are complementary, as they both highlight the similar characteristic of LAWS as being independent of the intervention by a human operator. Additionally, ICRC also fills the gap in DoD's definition by detailing the tasks performed by these systems.

In this chapter, I aim to clarify the above definitions of LAWS and analyze the notion of autonomy in machines in general. For that purpose, the chapter proceeds with an analysis of autonomy in machines. Then, I explore the oft-used framework of autonomy in weapon systems: the loop framework. Consequently, the chapter constitutes an essential starting point for the ethical discussion of autonomous weapon systems in the following chapters.

2.1 Autonomy in machines

To better grasp what autonomy means in weapon systems, it is beneficial to understand some fundamental aspects of autonomy in machines in general. The word autonomy comes from the combination of two Greek words: "autos," meaning "self," and "nomos," meaning "rule" or "law" (Smithers, 1997, p. 94). In a similar vein, Oxford Learner's Dictionaries (n.d.) describes autonomy as "the ability to act and make decisions without being controlled by anyone else." An entity, thus, possesses autonomy if it can act in accordance with its own rules and is not controlled by anyone else. Even though these considerations offer an understanding of autonomy as being independent of human intervention, they come short of providing a clarification of what autonomy means in machines. That is, there is a need for clarification on the autonomy component of these systems because independence from human operators, as will be apparent later in this chapter, is not sufficient to understand autonomy in machines. For this reason, in the rest of the chapter, I will analyze other parameters that provide a clearer understanding of machine autonomy. Then, I will discuss the programming features of the machines that allow them to operate autonomously, i.e., rule-based and machine-learning techniques.

2.1.1. Parameters of autonomy

In simple terms, an agent is autonomous if that agent "determines its actions for itself based only on its internal state,…that is, if the determination of the agent's behavior is local and without input from other agents" (Beavers & Hexmoor, 2004, p. 95). This description points out one of the characteristics of autonomy as being able to function independently of external control or intervention. Under this definition, a washing machine would be an autonomous entity. Once we plug the washing machine in and press the start button, it carries out operations such as water intake to its drum, detergent addition, spinning, and rinsing. It performs all these tasks or functions independently, i.e., autonomously. Similarly, self-driving cars, somewhat tautologically, are expected to drive on their own. They must obey the traffic rules, stop at red lights and cross at green lights, and give way to pedestrians at pedestrian crossings, etc.

Although capable of performing specific tasks independently (for instance, running a pre-defined washing cycle), washing machines can only perform a narrow range of functions. They cannot adapt to novel inputs or make complex decisions beyond their programmed instructions. As a result, they are not referred to as autonomous in the same sense as self-driving cars.

On the other hand, self-driving cars are designed to operate autonomously in dynamic and unstructured environments. They should process perceptual input from their environment and navigate complex traffic conditions independently of human operators. The difficulty of handling tasks in a complex and dynamic environment is why self-driving cars are considered autonomous, and their development requires more effort and time than simple household appliances like washing machines. We are more likely to categorize self-driving cars as autonomous but not washing machines primarily due to the difference in their respective complexity. The tasks washing machines perform are limited and explicitly pre-determined by their programmers. Within the selected program, washing machines will always generate the same results for the same inputs. In other words, they are highly predictable. This type of machine, i.e., machines operating relatively simple tasks within highly predictable and structured environments, is called automated (Beernaert, 2018; Heyns, 2013).

On the other hand, self-driving cars should be flexible to respond to different inputs depending on dynamic factors such as traffic, other vehicles, and pedestrians. In contrast to washing machines, what makes self-driving cars autonomous is "the ability… to deal with uncertainties in its operation environment" (Boulanin & Verbruggen, 2017, p. 6). Thus, operating independently alone is insufficient for categorizing machines as autonomous. To comprehend autonomy in machines, another factor must also be taken into account. This factor has been conceptualized in various ways in the literature. Some argue that complexity should be attributed to the machine itself. This leads them to conceptualize this factor as "the complexity of the machine" (Horowitz & Scharre, 2015, pp. 5-6) or "sophistication of the machine" (de Vries, 2023, p. 45).

Similarly, Ezenkwu and Starkey (2019) perceive autonomy as the property of the machine. Thus, they propose several attributes a machine should have to be categorized as autonomous. These attributes are "perception, actuation, learning, context-awareness, and decision-making" (Ezenkwu & Starkey, 2019, p. 2)[1]. As it may be apparent, Ezenkwu and Starkey (2019) particularly focus on the machine's complexity to assess its autonomy. These attributes refer to the abilities of the machine. For instance, perception is the machine's ability to process sensory input from its environment, and actuation is the ability to act upon the environment. Similarly, learning, context-awareness, and decision-making refer to the power of the machine to learn from the sensory input, adapt to the context in which it operates, and make decisions due to learning and adaptability. Others, on the other hand, place the complexity in the task assigned to the machine, and therefore, they refer to this

[1] Ezenkwu & Starkey categorize these attributes as low-level attributes. For them, an autonomous machine must have these attributes. However, they also discuss high-level attributes, which are subject to continuous research. High-level attributes are "domain-independence, self-motivation, self-recovery, and self- identification of goals" (2019, p. 2). They claim that these more advanced attributes are not must-haves but are subject to ongoing research. For this reason, I only discuss the must-have attributes of an autonomous machine as proposed by them.

parameter as “task complexity” (Beernaert et al., 2018, p. 2824). Alternatively, some discuss “environmental difficulty” as an additional parameter to the task complexity (Huang et al., 2004, p. 4). These conceptualizations might confuse a clear understanding of the complexity parameter because there are three seemingly distinct but highly intertwined factors: machine complexity, task complexity, and environmental difficulty. However, following Bradshaw et al. (2013), I think the complexity parameter consists of the interactions between these three factors. This means that the complexity shaping the autonomy of a machine depends on the three-dimensional interactions among the machine, the task assigned, and the environment or situation in which the machine is expected to operate. The complexities in these three factors shape the extent to which a machine is autonomous. For example, a structured and static environment is relatively simple and predictable, with fewer variables and less uncertainty. In this context, the tasks assigned to self-driving cars, such as maintaining a constant speed in an empty zone, do not require highly sophisticated algorithms and adaptation capabilities.

Contrary to structured environments, the task and the machine’s complexity increase significantly in unstructured and dynamic environments such as traffic-dense city centers. The self-driving car must have adaptation capabilities to execute its functions based on a multitude of unpredictable inputs, such as recognizing pedestrians, other vehicles, traffic lights, and so on. The machine would need more advanced algorithms and sensors to handle these tasks. In this scenario, the environment would impact both the machine’s and task’s complexity. Thus, the three-dimensional relation among the machine, the task, and the environment becomes essential for assessing autonomy. Consequently, the above-discussed three-dimensional complexity is a more comprehensive approach to determining autonomy than merely concentrating on the machine’s complexity. So far, this analysis demonstrates two parameters for assessing autonomy in machines. The first is human-machine interaction, more precisely, the independence from the human operator, and the second is the three-dimensional complexity in the machine, the task, and the environment.

Another parameter shaping the autonomy in machines is which type of decisions are automated in a system. When considering a machine, it is crucial to shift the focus from whether it is entirely autonomous to examining which specific decisions within the system are automated and which require input from human operators (Scharre & Horowitz, 2015). Classifying what critical functions make a machine autonomous when performed independently of an operator is important. Drawing a parallel with the washing machine example, automating the process of when the machine takes detergent to its drum carries a different level of difficulties and risks than automating the function of braking in traffic-dense and pedestrian-crowded urban areas. So, it becomes essential to determine what functions, when automated, make the system an autonomous system. For instance, if washing machines incorporate an additional function, such as automatically ordering detergent when a shortage is detected, would we classify them as autonomous? Similarly, would they meet the autonomy criteria if they could learn from our washing habits and prioritize items in the pile of dirty clothes based on usage patterns? Simply put, the essence of this factor is to show that what type of functions are automated plays a vital part when defining autonomy rather than classifying the system as a whole autonomous.

As a result, three parameters offer a comprehensive and more explicit approach to assessing machine autonomy rather than focusing solely on human-machine interaction regarding machine independence from a human operator. However, the parameters should not be viewed separately as if they exclude each other. Even though they can be analyzed individually, they are intertwined and influence each other. For example, the complexity parameter affects the extent of human involvement and oversight over the machine's operations. An increase in the ability of independent functioning of the machine may lead to an increase in the complexity and the type of functions automated.

2.1.2. Programming

Autonomous machines can execute tasks per their algorithmic rules, allowing them to operate without direct human control or intervention. Nevertheless, it is essential to consider the source and nature of these rules, as they contribute to shaping the boundaries of the machine's autonomy. For this purpose, we must examine the technologies or methods that enable machines to possess the ability to operate autonomously. Computer programs are instructions that process input data to produce an output (Alpaydın, 2016). In traditional programming, these instructions are defined by the programmer. Each operation step is carefully written in a specific order, and the input and output sets are pre-determined.

Several outcomes can occur when a program receives input that deviates from the expected parameters. In some cases, the system may encounter an error and crash, unable to process the input within the constraints of the existing algorithm. Alternatively, it may generate an error message, notifying the user that the input is invalid and assisting the user to enter the correct input.

For example, consider a situation where you are asked to provide your date of birth in the format of ‘month/day/year’ to buy an online train ticket where you can have a discount if you fall under a particular age group. Now, imagine that you mistakenly enter your date of birth in the format of ‘day/month/year,’ as it is generally used in your country.

In this example, your input can result in various issues. The program may inaccurately calculate your age, so you may not benefit from the discount. If your day of birth falls on or after the 13th of the month, the program might issue an error message to you, notifying you that only numeric inputs between 1 and 12 are permissible for the month component of the birth date.

The workings of such programs employ conditional logic, so specific conditions often need to be met for the computation to proceed accurately. In our example, the program checks:

> IF the input is less than or equal to 12, THEN the program continues with the calculation.
> ELSE (i.e., if the input is not between 1 and 12), the program generates an error message.

In algorithms with this type of programming feature, since the programmer determines how the program will execute its operations, it is possible to anticipate the specific outputs generated for each input provided. Given sufficient time, humans can replicate the execution of the program step by step, consistently achieving the same result for the same input.

This deterministic nature of programming allows for predictable and stable outcomes. However, there are scenarios where this programming type may be inappropriate for a given task. For example, if the programmers themselves do not know how to write rules for the execution of a task, then rule-based programming is not efficient to employ in such scenarios.

2.1.3. Machine learning

Computers have the potential to perform a wide range of tasks as long as we can define the operations to be executed accurately. However, it can be challenging to accurately represent the operations for some tasks because as the complexity of the task increases, so do the complexities of space, time, and human elements (Domingos, 2015).

As a task becomes more complex, a greater amount of data is required to be stored in the program's memory. Likewise, the processing time needed to perform the task also increases, so this would cause a need for more powerful computational resources. Moreover, as algorithms become more elaborate, it becomes harder for humans to comprehend the interactions between different parts of the algorithm. This leaves programmers subject to failure in fixing errors in algorithms, and even a "[o]ne tiny error in an algorithm" may result in the explosion of a rocket or electric cut for millions (Domingos, 2015, p. 40). Also, for some complex tasks, a programmer might not know how to define specific functions for a given task, so it becomes essential to employ the methods by which the complex tasks can be executed.

AI, specifically machine learning (ML) algorithms, are employed to overcome these challenges. AI is an umbrella term for computational programs capable of displaying near-human-like cognitive abilities. ML is, on the other hand, a sub-category of AI. On the contrary to traditional programming, ML and AI generally offer more efficient ways to cope with complex tasks. Self-driving cars exemplify how ML algorithms work for complicated tasks. Engineers did not write strict rules for every particular action the vehicle would take. Instead, the car stays on the road by learning from the driver's behavior. This process involves ML algorithms that enable the vehicle to adapt to its environment to process novel inputs that cannot be programmed beforehand (Domingos, 2015).

As discussed earlier, humans provide detailed instructions to the computer in traditional programming. In machine learning, however, instead of giving explicit instructions to the computer, it is fed with sample data, also called training data. After being fed with data, the program builds a model from the data and then processes new inputs following the model built from the training data (Alpaydın, 2016). This method is called machine learning because the machine learns from the sample data. The data mostly replaces the role of the programmer. For example, in supervised learning, the programmer provides the machine with sample data, i.e., giving sample inputs and desired outputs. However, the programmer no longer explicitly defines the instructions to process an input and its corresponding output. Therefore, what is required for machine learning is not elaborate algorithms but vast amounts of sample data that the machine will learn how to execute its tasks (Alpaydın, 2016).

For example, let us imagine a program that classifies texts based on different eras in the history of philosophy, such as Ancient Greek, medieval, modern, etc. We provide sample texts labeled following their corresponding period. These texts are the training data for the program. The program, then, learns from these texts. For instance, it analyzes the syntax, philosophical concepts, and tones and then comes up with patterns for each era in the history of philosophy. Thus, the program builds a model to categorize new texts not in its initial sample data. However, one could ask why there is a need for machine learning when the classification of the texts can be done with traditional programming. We could write specific algorithms that would classify texts based on their eras by detailed instructions and criteria for a specific era.

Even though it is possible that traditional programming would accomplish this task, some tasks cannot be done with traditional programming. For example, self-driving cars or autonomous vehicles, where vast possibilities they may encounter, make it impossible to anticipate and pre-program instructions for every situation. Self-driving cars should recognize the traffic lights' position, the lights' colors, the time to pass, the position of the other vehicles and pedestrians, and the time to stop or not stop for pedestrians. Moreover, it should do so in a dynamic environment because these variables constantly change through time. Therefore, a programmer cannot predict every scenario and write algorithms for these scenarios.

Self-driving cars collect data from sensors, cameras, GPS (global positioning system), lidar (light detection and ranging), etc. In the training stage, the human operators drive the car to collect data, which will be used later for self-driving mode. The learning component of machine learning means that the car learns how to adapt and respond to dynamic surroundings and changing variables from its training. By machine learning, the task of driving can be automated, while it is virtually impossible to do so by writing every instruction the car will execute during its driving time. However, the capability of adapting to changing environments also means that there may be instances where the outputs are not entirely predictable, which leads to safety concerns. Therefore, the ability of machine learning to handle complex tasks and adapt to changing environments while processing new inputs has advantages and disadvantages. On the one hand, it allows machines to improve their performance over time and automates complicated tasks in various fields. However, on the other hand, this ability comes with novel risks and challenges, such as the need to anticipate how the machine will perform in a changing and unforeseen environment.

2.2. Autonomy in weapon systems

I have analyzed the programming background of autonomy in machines and how these methods allow a machine to be categorized as autonomous. The technical analysis of traditional programming and machine learning will be important to examine the ethical implications for LAWS. However, before jumping on the ethical concerns, I will discuss another prominent framework used in the literature to understand better what autonomy is within the context of LAWS.

2.2.1. The loop framework

In addition to the analyses in the previous sections, there is a framework commonly used in the literature when discussing the nature of autonomy in weapon systems. This framework is the loop framework, which originally describes the targeting process in the military context.

Boyd, a fighter pilot, first developed the original loop framework for assessing the cycle a fighter pilot completes when targeting an enemy jet (Anderson & Waxman, 2013). The cycle is known as the OODA loop, standing for observation, orientation, decision, and action. Osinga briefly describes the cycle as follows:

> [O]bservation is sensing yourself and the world around you. The second element, orientation, is the complex set of filters of genetic heritage, cultural predispositions, personal experience, and knowledge. The third is decision, a review of alternative courses of action and the selection of the preferred course as a hypothesis to be tested. The final element is action, the testing of the decision selected by implementation (Osinga, 2005, pp. 2-3).

By carrying out this four-element loop, a targeting takes place in the military context. However, it is also important to undergo this loop as fast as possible because the party, quicker and more accurate than the opponent, will have a military advantage over the enemy (Anderson & Waxman, 2013). Compared to humans, who may have limitations in terms of response time, automation of the loop can offer advantages such as faster response time, more accuracy, and efficiency while undergoing the loop. These advantages over humans when completing the loop are among the driving forces to develop autonomous weapon systems because they will leave the enemy in reactive mode (Schmitt & Thurnher, 2013), and if "other things being equal, the faster system wins the engagement" (as cited in Anderson & Waxman, 2017, p.1102).

The loop framework is appropriated to refer to the human operator's roles in the autonomy discussion of LAWS. This framework also corresponds to one of the dimensions previously discussed about autonomy in machines: human-machine interaction. The loop framework details the human-machine relationship; rather than simply focusing on whether a human operator is present or not, it deals with the levels of autonomy in weapon systems and the human control over the system's operations. According to the loop framework, human operators play three different roles in relation to the machine: "human-in-the-loop," "on-the-loop", and "out-of-the-loop" (Scharre & Horowitz, 2015, p. 8).

First is *human-in-the-loop*, which indicates that the weapon system can perform the targeting loop only if the operator actively selects the targets to be engaged. Human-in-the-loop weapon systems are also known as semi-autonomous weapon systems. The operator maintains control over the decision to target selection, but the systems can autonomously perform functions such as "acquiring, tracking, and identifying potential targets" (DoD, 2023, p. 23). The examples of weapon systems that fall under this category are fire-and-forget munitions, where the system optimizes the precision of engaging the target the operator selects. The human operator chooses the target to be engaged and, after launch, munitions correct their direction with onboard sensors to "home in on moving targets" (Scharre & Horowitz, 2015, p. 9). Put analogically, the human operator is responsible for pulling the trigger, but the bullets employ gadgets such as sensors to improve the probability of hitting the enemy.

Second is *human-on-the-loop* systems, also called human-supervised or operator-supervised autonomous weapon systems. In such systems, the operator's role is to monitor the system's activity rather than selecting with appropriate knowledge the course of action the system will carry out. The DoD defines these systems as "designed to provide operators with the ability to intervene and terminate engagements, including in the event of a weapon system failure before unacceptable levels of damage occur" (DoD, 2023, p. 22). If the operator does not intervene, however, the system will perform the tasks in the loop independently (Scharre, 2020). An example of a human-on-the-loop weapon system is Israel's Harpy, developed and promoted as an "autonomous weapon for all weather" by Israel Aerospace Industries (IAI). The Harpy, after launch, loiters a predefined area to search for enemy radars. Upon detection, it hits and destroys the radars. IAI describes Harpy's performance as being able to conduct "autonomous operation," and the operator has the supervisory ability to "abort attack in case of target shut down." (IAI, n.d.).

The final, third role is *human-out-of-the-loop*. This type of system reflects the general image of LAWS the most. It receives significant attention because the system, in this case, acts independently without any control as in human-in-loop systems, or supervision by an operator as in human-on-the-loop systems. DoD includes operator-supervised (on-the-loop) weapon systems in this category but also notes that they are not "limited to operator-supervised autonomous weapon systems that are designed to allow operators to override operation of the weapon system" (DoD, 2023, p.21). This means that other systems that can engage targets without the operator's ability to intervene or halt the operation are included. However, human-on-the-loop systems would also be examples of human-out-of-the-loop systems because they may be deployed in fully autonomous mode. One significant example would be the robotic sentry system, Super aEgis II, developed by South Korean company DODAAM. This system can operate in all three modes, i.e., humans make firing decisions, humans can halt the system’s actions, and the system operates entirely autonomously (Boulanin & Verbruggen, 2017). In autonomous mode, Super aEgis II uses thermal sensors and cameras to detect the heat and motion of a potential target and, based on this input data, makes targeting decisions without human control or supervision. Currently, sentry systems can be employed in demilitarized zones, where if a human is detected, then it would be a legitimate target. Thus, sentry systems in fully autonomous mode cannot be used in areas other than demilitarized zones because there is as yet no software system to detect if the target is a civilian, combatant, surrendering, etc.

Within the literature, there are ethical and legal challenges surrounding all three types of systems, but the locus of the debate revolves around human out-of-the-loop systems. For the present thesis, my analysis will also concern these systems.

In summary, human-in-the-loop systems leave the decision to kill a specific target to a human operator. In human-on-the-loop systems, the entire cycle of targeting, including the kill decision, is done by the weapon itself, but the human remains in supervisory control over the system's operation; also, human operators can override the system’s operation. Unlike these two systems, human-out-of-the-loop systems will not require an operator once deployed, so they differ in their ability to undergo the entire cycle of targeting without an operator’s direct control or supervision.

While the loop framework is prominent in the literature to define LAWS, there are problems pertaining to it. The loop framework falls short of encompassing the autonomy of these systems. The framework solely focuses on the two parameters of autonomy discussed previously, i.e., the role of human operators in the decision-making of the system and the type of functions automated. However, it fails to offer an understanding of the complexity parameter in the autonomy of LAWS, that is, the complexity of the machine, the tasks, and the environment. Taddeo & Blanchard (2022b) point to a similar lack in the definitions of LAWS proposed by the states and NGOs. They analyzed 12 definitions from states and international organizations such as the US, the UK, China, and NATO. They found out that “only the French and the Chinese definitions stress the adapting capabilities, specifically the definitions mention machine learning capabilities of [L]AWS as a key characteristic” (Taddeo & Blanchard, 2022b, p. 37). The complexity of the technological characteristics is also an issue for the loop framework since it heavily focuses on the role of humans and the functions executed by the machine. However, machine learning is important when considering the ethical and legal problems pertaining to the use of LAWS. As discussed earlier, machine learning is a useful application in complicated tasks. While some tasks can be automated through rule-based programming, some tasks, such as driving, are considerably difficult to define each action through rule-based algorithms. The militaries' interest in AI also proves that rule-based algorithms are "being replaced by AI-based systems" (Taddeo & Blanchard, 2022b, p. 37). ICRC, similarly, points out that the "future developments could include increasing *adaptability* [emphasis added] of these weapon systems to their environment" (ICRC, 2016, p. 2). For this reason, while discussing the ethical and legal implications of LAWS, it is also essential to remember that these machines' complex abilities play a vital role in that discussion.

Throughout the thesis, I will employ the technological aspect of LAWS and the concept of human-out-of-the-loop. Thus, when referring to LAWS, I will refer to a weapon system that can autonomously undergo the targeting cycle without human intervention, relying on its technological capabilities. The capabilities that allow a system to operate in human-out-of-the-loop mode is an important factor to consider in the ethical discussions of these systems. Failing to consider such capabilities has led to disagreements in the ethical aspects of LAWS. In the next chapter, technological capabilities will be important as they play an essential role in shaping the perspectives of those either in favor of or against the use of LAWS. Another clarification needed before moving on to the next chapter is that the use of the term "lethal autonomous weapon systems (LAWS)" in this thesis is deliberate, as it emphasizes what I consider the most critical aspect: the function of the actual killing. While different tasks within the targeting loop may present distinct problems associated with them, I believe that lethality is an important aspect that should be emphasized in naming these systems.

CHAPTER 3

ETHICAL ISSUES SURROUNDING LAWS

Under the umbrella organization Campaign to Stop Killer Robots, more than two hundred international, regional, and national non-governmental organizations (NGOs), including the International Committee for Robots Arms Control, Human Rights Watch, Amnesty International, Future of Life Institute and over 90 states from different parts of the world call for a legal treaty that will "create the prohibitions and regulations that will ensure human control" over LAWS (Stop Killer Robots, n.d.). Some high-tech companies working in AI and robotics, such as Tesla, Google DeepMind, and Clearpath Robotics, and individuals such as Max Tegmark, Stuart J. Russell, and Elon Musk signed an open letter to urge the UN to fasten the process of "strong international norms, regulations and laws against lethal autonomous weapons" (Future of Life Institute, 2018).

In another open letter, which is also endorsed by the philosophers Daniel C. Dennett and Noam Chomsky, roboticists claim that the development of LAWS will reduce the threshold of raging war because switching human soldiers with machines will lessen the risks of soldier casualties to rage a war for the party owning LAWS. Also, illegal organizations and terrorists may easily access the required material resources to mass-produce such machines. They also warn against the possible use of LAWS to kill particular ethnic groups, increasing genocides (Future of Life Institute, 2016). For these reasons, high-tech companies and roboticists declared that they will not take part in the research and development of LAWS. However, there are "countless university laboratories… and commercial enterprises" working in the LAWS-related technologies (Altmann & Sauer, 2017, p. 125). To discuss the above-mentioned concerns, the first CCW Group of Governmental Experts meeting on discussing the risks and challenges of LAWS was held in 2014.

The most recent sessions at CCW on LAWS were held on 15-19 May 2023. However, after a decade of discussions, the UN Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems ended with a conclusion that International Humanitarian Law (IHL) fully applies to the emerging technologies in LAWS. This means that contrary to the suggestions of the campaigners for stopping the use and development of LAWS, there is as yet no formal regulation or prohibition mandated by the Convention on the issues related to the use or development of LAWS.

In this chapter of the thesis, I will discuss the ethical and legal concerns surrounding LAWS. I will first analyze how a ban on particular weapons is achieved. After, I will discuss the arguments in favor of and against the deployment of LAWS in relation to *jus in bello* principles in warfare.

3.1. How to ban weapons?

The use of many weapons is banned on the battlefield. If you have been a part of or seen footage of protests in many countries, you have been directly exposed to tear gas or seen that law enforcement uses tear gas against protesters. However, the use of tear gas is banned on the battlefield because its use falls under the Chemical Weapons Convention.[2]

The prohibition and regulation of certain weapons are often achieved through legal treaties and conventions that aim to restrict their use. While chemical weapons are a notable example, there are various other weapons that have been subject to such agreements. There are bans on anti-personnel mines[3], blinding lasers[4], cluster munitions[5], bacteriological(biological) and toxin weapons[6]. There is a treaty on the prohibition of nuclear weapons[7], but it is worth noting that major states possessing

[2] ICRC, https://ihl-databases.icrc.org/en/ihl-treaties/cwc-1993?activeTab=default

[3] ICRC, https://ihl-databases.icrc.org/en/ihl-treaties/apmbc?activeTab=default

[4] ICRC, https://ihl-databases.icrc.org/en/ihl-treaties/ccw-protocol-iv?activeTab=default

[5] ICRC, https://ihl-databases.icrc.org/en/ihl-treaties/ccm-2008?activeTab=default

[6] ICRC, https://ihl-databases.icrc.org/en/ihl-treaties/bwc-1972?activeTab=default

[7] ICRC, https://ihl-databases.icrc.org/en/ihl-treaties/tpnw-2017?activeTab=default

nuclear weapons have not signed this treaty. Among NATO members, the Netherlands is currently the only signatory of the treaty.

The restriction or prohibition of specific types of weapons is not a new concept and has been part of international discussions for centuries. In fact, the first recorded international ban on a weapon goes back to 1675. Strasbourg agreement was signed by France and the Holy Roman Empire in 1675 to prohibit the use of poison bullets (Organisation for the Prohibition of Chemical Weapons, n.d.).

Strasbourg agreement is a bilateral agreement bounding France and the Empire, but a multilateral international treaty that aims to restrict or prohibit the use of certain types of weapons can also be established. The UN Convention on Certain Conventional Weapons (CCW), formally known as the 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, is a multilateral international treaty. The CCW provides a platform for the member states to discuss and take necessary formal actions to regulate weapons that are considered to impose unjustified suffering on combatants or cause indiscriminate damage to civilians and nonmilitary objects. For example, anti-personnel land mines are banned on the grounds that they have an indiscriminate effect on non-combatants. The impact of landmines can extend beyond the period of active conflict. This causes problems for post-conflict recovery efforts. After the armed conflict, unexploded land mines can directly harm and cause fatalities to individuals who unknowingly trigger them. Moreover, landmines may contaminate the lands or impede construction of infrastructure projects such as houses, roads, and so on. (Melzer, 2016). As a result, landmines are considered to impose damage on civilians and civilian objects and, therefore, are banned. Another regulatory instrument specifically referring to the use and development of new weapons is Article 36 of Additional Protocol I to the Geneva Conventions (ICRC, AP I, Art. 36). Article 36 requires states to conduct legal reviews of new weapons or means of warfare to ensure their compliance to the IHL, and it is an obligation which "applies to *all States irrespective of their treaty obligations* [emphasis added] because they are legally responsible for ensuring that they do not use prohibited weapons or use lawful weapons in a manner that is prohibited" (Melzer, 2016, p.122).

Thus, the Geneva Conventions and, more specifically, Article 36 of Additional Protocol I provide a conceptual framework for the Convention to regulate and, if necessary, prohibit the use and development of certain weapons. The companies, NGOs, and academics who oppose the use and development of LAWS are advocating for the creation of a legally binding treaty under the Convention, which explicitly regulates and prohibits the use and development of LAWS.

3.1.1. International humanitarian law

Since it is a significant element in the debate on LAWS, the rules governing the conduct of warfare should be discussed to better grasp whether LAWS pose problems in complying with these rules and whether they should be subject to restrictions or an outright ban. International Humanitarian Law (IHL), also known as the law of armed conflict or laws of war, is the body of law that governs the conduct of warfare. When we think of war, we may tend to think of the horrors or destructive nature of it. However, even in times of war, there are rules that must be obeyed to limit the sufferings of those affected by the horrors of war.

IHL is the umbrella title for the varying sources of the rules on the battlefield. The main body of the rules is based on the Geneva Conventions and their Additional Protocols, which set the rules for the conduct of warfare. These rules not only limit the harm inflicted on civilians but also regulate how combatants who are no longer participants of the war, such as wounded, sick, and prisoners of war, should be treated.

Both treaty law and customary law form the foundations of IHL. The difference between the two is that treaty IHL is the Geneva Conventions and their Additional Protocols, and they "are written conventions in which States formally establish certain rules" (ICRC, n.d.). Customary IHL, on the other hand, bound all states, and its rules come from "general practice accepted as law" (ICRC, n.d.). The general practices include examples such as "military manuals, national legislation,… instructions to armed and security forces, comments by governments on draft treaties,… statements in international organisations [etc.]" (Henckaerts & Doswald-Beck, 2009, p. XXXVIII). The treaty IHL and customary IHL are interconnected and not mutually exclusive. While some rules of customary IHL are explicitly written and codified into the treaty IHL, customary IHL goes beyond treaty obligations. Even states not party to specific treaties are still bound by customary IHL. For example, although the United States is not a party to Additional Protocol I to the Geneva Conventions, it conducts legal reviews of new weapons as required by Article 36 of Additional Protocol I (Anderson & Waxman, 2013). The provisions of both treaty and customary IHL help minimize the destructive effects of war and prohibit the inhumane treatment of those affected, regardless of their status as combatants or non-combatants.

3.2. LAWS and principles of *jus in bello*

Just war theory is the ethical framework that provides criteria for when to go to war (*jus ad bellum)* and for ethically acceptable conduct in warfare (*jus in bello*). Going to war is justified if the six criteria are met: "just cause, proportionality, necessity, last resort, right authority, and reasonable likelihood of success" (Leveringhaus, 2016, p. 12). On the other hand, ethical conduct in warfare consists of three criteria: "distinction, proportionality of means, and necessity" (Leveringhaus, 2016, p. 12). *Jus in bello* principles are related to the debate of LAWS, as *jus ad bellum* concerns questions such as whether it is ever justified to go to war. *Jus in bello,* on the other hand, concerns the questions of who is a legitimate target, what means are justified means in warfare, etc. Thus, the ethical discussion mostly focuses on whether LAWS can comply with the principles of *jus in bello*.

Two fundamental principles of *jus in bello* and IHL are considered the most related to LAWS: the "principle of distinction and proportionality" (Asaro, 2012, p. 688). The distinction principle is "one of two principles in the law of armed conflict recognized as 'cardinal' by the International Court of Justice" (Schmitt & Thurnher, 2013, p. 251). This principle legally enforces the parties of the conflict to "distinguish between the civilian population and combatants and between civilian objects and military objectives and … direct their operations only against military objectives" (ICRC, AP I, Art. 48). This principle also entails that new weapons or methods of attack are prohibited if they cannot be targeted exclusively towards military personnel or military materials (ICRC, AP I, Art. 51).

Cluster munitions, for example, are classified as indiscriminate weapons due to their design. Cluster munitions are explosive weapons that release a group of individual munitions simultaneously, covering a wide area. However, the individual munitions cannot be targeted toward specific enemy soldiers or objects. Their inherent limitation of being unable to target specific enemy soldiers or objects makes them fall into the category of indiscriminate weapons. As a result of this, cluster munitions are prohibited under the Convention on Cluster Munitions.

Another example that illustrates the violation of the principle of distinction is the usage of artillery in certain areas, such as densely civilian-populated areas. Artillery, long-rage weapons intended to target military objectives from afar, becomes indiscriminate when deployed in regions with high concentrations of civilians. Despite being a legal weapon, employing artillery in such scenarios would contravene the principle of distinction, as it might fail to discriminate between combatants and non-combatants. These two examples demonstrate two cases for the principle of distinction. The former involves the use of an inherently indiscriminate weapon, and the latter shows a situation where the indiscriminate use of an otherwise legal weapon violates the principle.

Critics of lethal autonomous weapon systems (LAWS) claim that LAWS are highly likely to be incapable of complying with the principle of distinction (Asaro, 2012; Human Rights Watch, 2012; Sharkey, 2012; Sparrow, 2016). For Sparrow, these systems cannot distinguish combatants from non-combatants because they are "somewhat unpredictable" (2007, p. 65). Similarly, Asaro argues that, compared to human intelligence, LAWS "will have only highly limited capabilities for learning and adaptation at best, it will be difficult or impossible to design systems capable of dealing with the fog and friction of war" (2012, p. 692). Asaro (2012) further claims that the complexity of the battlefield exceeds the anticipations of military roboticists, particularly in terms of the ability of these machines to comply with the principle of distinction. Noel Sharkey, a computer scientist and spokesperson for Stop Killer Robots, agrees with this sentiment by emphasizing three aspects required for the principle of distinction. First, he claims that these systems lack sufficient "sensory or vision processing systems" to distinguish combatants from civilians (Sharkey, 2017, p. 179). Second, the lack of a clear definition of civilian is in itself a challenge for any attempt to code a program to distinguish civilians because civilians are defined as "someone who is not combatant" in IHL (Sharkey, 2017, p. 179; Sharkey, 2012, p. 789). Lastly, Sharkey argues that human interpretative judgment is necessary for ensuring compliance with the principle of distinction because vision processing is insufficient for making distinction decisions. That is, even if sensory technologies improve to an advanced level, machines still would not have "battlefield awareness or common sense reasoning to assist in discrimination decisions" (Sharkey, 2017, p. 179).

To evaluate the claims made by Sharkey, examining the conditions that determine whether an individual is classified as a combatant or non-combatant and how such statuses are assigned in warfare is beneficial. When determining whether a person is a combatant or a civilian, wearing a uniform is often considered as one of the initial conditions. Uniforms help identify individuals as members of an armed force. However, under some circumstances, even individuals wearing uniforms are considered "hors de combat", meaning they are no longer legitimate targets. Combatants who are unable to continue participating in hostilities due to "wounds or sickness", or "*clearly* [emphasis added] expresses an intention to surrender" cannot be targeted as legitimate military objectives (ICRC, AP1, Art. 41). Detecting the *intention* to surrender poses an even more significant challenge, as the ways in which a soldier may clearly express their intention to surrender can vary significantly, such as displaying a white flag, verbal communication, raising both arms or laying the guns down etc. Consequently, in such situations, even sensory and vision processing systems to detect whether a person is wearing a specific uniform would not be sufficient for engaging with that target. The system should also detect if the uniform-wearing person has hors de combat status due to sickness, wounds, or intending to surrender.

Another problem with determining combatants wearing a uniform is the nature of contemporary warfare. Given the evolving nature of contemporary armed conflicts, uniform-wearing criteria may only be sufficient in some situations. For example, in non-international conflicts where a state engages in an armed conflict with a militia, distinguishing between combatants and civilians becomes particularly challenging, as rebel groups may not wear a distinctive uniform. One could argue that carrying arms, such as a rifle, would be sufficient to identify and target that person. For instance, sensory systems can detect rifles or military equipment to select military targets. Yet, the challenges in international armed conflicts still apply in these cases, meaning that the system should also detect if the person carrying a rifle is already wounded, unconscious, or surrendering and, therefore, not actively engaged in the war. Also, the machine should "recognise insurgents burying their dead, or children being forced to carry rifles" (Sharkey, 2019, p. 76).

Given these points, compliance with the principle of distinction poses a two-fold challenge. First, it requires detection of whether a person is an enemy soldier or non-combatant. Moreover, if identified as a soldier, one should detect if the target holds the status of hors de combat. For the critics of LAWS, these challenges cannot be overcome through algorithms (Asaro, 2012; Sharkey, 2017).

Accordingly, critics contend that LAWS will likely fall short of complying with the principle of proportionality in *jus in bello*. The principle of distinction is bounded by another "cardinal" principle that regulates the harm to civilians or civilian property that cannot be avoided in armed conflict: the principle of proportionality. IHL prohibits attacks "which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be *excessive* [emphasis added] in relation to the concrete and direct military advantage anticipated" (ICRC, AP I, Art. 51). Thus, it is not legitimate to target civilians or civilian objects intentionally. However, if such targeting cannot be avoided when attacking military objectives, they are tolerated as collateral damage or side-effects of the intentional attacks on the military targets. The principle of proportionality mandates parties to a conflict not to cause excessive collateral damage as a side-effect of the military advantage to be gained. Sharkey (2017) argues that there are two kinds of proportionality: easy and hard proportionality. According to him, LAWS can only function with easy proportionality; that is, machines can help to reduce "collateral damage by choosing the most appropriate weapon or munition and directing it appropriately" (Sharkey, 2017, p. 179). For example, precision-guided munitions have reduced indiscriminate and disproportionate attacks by enabling more accurate targeting with their onboard sensors. In a similar vein, LAWS could help reduce disproportionate attacks by employing advanced calculation skills to assess the potential risks associated with different courses of action and then selecting the most appropriate munitions to minimize collateral damage while maintaining military advantage. However, Sharkey adds that machines cannot make hard proportionality decisions, that is, to decide whether to apply lethal force for the military advantage to be gained “in the first place” (2017, p. 179). Sharkey thinks that civilian casualty and military advantage require “human qualitative and subjective decision about what is proportional to direct military advantage” (2017, p. 180). Thus, easy proportionality might be possible by employing LAWS on the battlefield and minimizing collateral damage of an intended attack. However, the decision to apply lethal force in a complex proportionality situation will remain challenging for machines to compute. They cannot calculate because proportionality is not a mere numerical calculation, unlike its potential connotations. Famous ethical dilemmas such as the trolley problem are, to some extent, examples of proportionality considerations. For example, how many children can be sacrificed for a high-ranking enemy leader is a proportionality problem. A machine may calculate the most accurate course of action with the least possible number of collateral damage. However, it cannot calculate whether any lethal force should be applied to the desired target in the first place. This argument is further supported by the practice of proportionality decisions in some situations where “[the] sensitive proportionality calculations…require, at least as a matter of policy in a democracy like the United States, an elected official or other senior political official to make the ultimate decision” (Beard, 2018, p.10). Given these points, critics argue that LAWS may neither be able to discriminate combatants from non-combatants nor calculate how much collateral damage is acceptable with regard to the concrete military advantage. As a result, they may leave “behind them a hecatomb of innocent victims” (Birnbacher, 2016, p. 118). These concerns raised by the opponents of LAWS stem from the inability to foresee how LAWS may operate in unstructured, complex environments, where there are unanticipated circumstances and a plethora of inputs to be processed, such as enemy behavior, changes in weather conditions, etc. This unpredictability aspect drives criticisms regarding the compliance of LAWS with the principles of distinction and proportionality.

3.3. Human, more-than-human: Advantages of LAWS

Proponents, on the other hand, claim that the unpredictability of LAWS cannot be a legitimate reason to ban these weapon systems since human combatants and existing military technologies also suffer from unpredictability on the battlefield. In fact, they argue, LAWS will reduce the unpredictability in warfare, and accordingly, they might reduce the number of war crimes, collateral damage, and civilian casualties (Arkin, 2009; 2010). Ronald Arkin, a prominent roboticist and roboethicist working in the field of military robotics, claims,

> the primary goal [of the development of LAWS] remains to enforce international humanitarian law… on the battlefield in a manner that is believed achievable, by creating a class of robots that not only comply with the restrictions of international law, but in fact outperform human soldiers in their ethical capacity under comparable circumstances (Arkin, 2010, p. 339).

Arkin (2009) proposes the ways in which these machines can overcome their human counterparts. He believes that these machines offer several advantages in warfare. Firstly, machines are not concerned with self-preservation and can act "selflessly" when necessary. Secondly, their sensors provide more precise "observations than humans currently possess" (Arkin, 2009, p. 29). Related to the machine sensory capabilities, LAWS can rapidly process more information from various sources before taking lethal action, outperforming what a human could do in real time. Thirdly, LAWS can be programmed without human emotions that might impair their decision and lead to acts of anger, vengeance, etc. Lastly, LAWS not only have the potential to act more ethically than human soldiers, but they can also possess the "capability of independently and objectively monitoring ethical behavior in the battlefield by all parties and reporting infractions," Arkin adds, "this presence alone might possibly lead a reduction in human ethical infractions" (2009, pp. 29-30).

To support his claims on the vulnerabilities of human soldiers, Arkin (2009) summarizes a report on Iraqi Freedom by the US Surgeons Generals Office and claims one of the findings of the report is that "soldiers that have high levels of anger… were nearly twice as likely to mistreat noncombatants as those who had low levels of anger" (p. 31). Considering the attitude of soldiers reporting ethical infractions in warfare, "45% percent of soldiers and 60% of marines did not agree that they would report a fellow soldier/marine if he had injured or killed an innocent noncombatant" (Arkin, 2009, p. 32). In addition to psychological limitations such as emotions, extreme weather conditions like fog, heat, cold, rain, snow, or sunlight can impair human judgment and lead to incorrect information or illegitimate targeting. Furthermore, hunger and thirst can impact human soldiers' actions on the battlefield. By developing robots that are less prone to these factors, the risk of illegitimate actions may be reduced.

Opponents of LAWS might conceive that Arkin's proposal for more ethical behavior by algorithms is unattainable. However, Schmitt & Turnher (2013) believe that not the proponents but the arguments of the opponents of LAWS are already "counter-factual" since military technologies have "advanced well beyond simply being able to spot an individual or object" (2013, p. 247). They argue that modern sensory systems can "assess the shape and size of objects, determine their speed,… listen to the object and its environs, and intercept associated communications or other electronic emissions" (Schmitt & Turnher, 2013, p. 247). LAWS may also interact with other systems to "monitor a potential target for extended periods in order to gather information that will enhance the reliability of identification" (Schmitt & Turnher, 2013, p. 247). These observations are accurate in some of the current weapon systems that are anti-material weapon systems, i.e., these systems are designed to target military objects. For instance, target recognition systems can detect a target based on the shape and height of tanks, speed and radio emissions of missiles and radars, auditory signals of submarines, etc. (Boulanin & Verbruggen, 2017). However, it should be noted that they still cannot detect if civilians are near these objects, thus rendering a possible attack illegitimate. These technological advances might serve as evidence for predictions that the technology required for developing LAWS to target humans and comply with *jus in bello* may not be in the distant future. Moreover, once developed, such machines might even be morally desirable, considering the ethical advantages they would have compared to humans. For achieving these prospects in the context of human targets, however, it should be justified that these machines can indeed abide by the principles of international humanitarian law (IHL) while minimizing the unpredictability human soldiers pose on the battlefield.

Arkin claims that LAWS will be constrained by strict rules derived from "Laws of War as well as the Rules of Engagement" (Arkin, 2009, p. 38). Thus, when faced with a decision, the machine will apply all the relevant constraints from the laws of war and then decide on the course of action. The actions of the machine will be limited by "a complex IF … THEN statement," and when all the conditions are met "in the IF part of that statement, [L]AWS engage their target" (Klincewicz, 2015, p. 164). Therefore, LAWS will only engage with targets only if the ethical conditions are met, which are constrained by rule-based if/then algorithmic structure.

Arkin also denies that, for ethical behavior, LAWS would require machine learning methods because he believes that the laws of war already provide general rules to implement to the machine “without the limitations and dangers of *training* [emphasis added]” (2009, p. 107). He acknowledges, for example, that “neural networks” would result in the loss of transparency where “the system cannot justify its decisions in any meaningful way; that is, explanations and arguments are not capable of being generated” (2009, p. 108).

Rule-based programming offers more predictable outcomes compared to other machine learning methods. This is because the deterministic nature of the if/then/else logical format significantly constrains the machine’s actions. It is, therefore, understandable that the proponents of LAWS resort to the rule-based method to overcome the challenges of unpredictability in the machine learning methods. Rule-based programming also influences the understanding of autonomy in these machines because it paints a picture that the machine is simply following human orders. McFarland claims that “sophisticated weapon systems are merely machines which execute instructions encoded in software, and… future highly autonomous systems envisioned by designers will not be anything more than that” (2015, p. 1326). He further exemplifies the workings of a potential autonomous weapon as:

> [I]f <camera image matches image in database> then <aim and fire> else <keep searching> [this] would make it appear that the UAV[unmanned aerial vehicle] itself is selecting targets when *actually the targets and the conditions under which they would be attacked were selected in advance by the system developers* [emphasis added] (2020, p. 34)

This example also shows an autonomous weapon with the rule-based algorithm for its operation, meaning that the machine merely executes the if/then/else functions. Unlike Arkin, McFarland does not rule out the possible use of machine learning approaches because the operation environment might require the machine to deal with novel inputs not foreseen by the designers. However, he contends that this would “not change the fact that the computer is only executing instructions formulated by its developer,” for him, the difference between a learner and rule-based algorithms only lies in the fact that the programmer of a learning system “writes a program the function of which is to formulate some optimum set of actions to be performed in response to environmental stimuli encountered during a mission” (2015, p. 1328). In other words, the programmer of a learning algorithm sets the goal of the machine in advance, so what the machine does is not more than finding the proper set of instructions to execute the functions for that goal, which “would otherwise have been issued to the machine directly by a human operator” (McFarland, 2020, p. 48).

While rule-based systems may be more predictable than other learning systems, the very nature of rule-based systems makes them unsuitable candidates for future lethal autonomous systems. For example, face recognition cannot be done with expert systems, i.e., by writing every rule for the machine. However, there has been significant progress in face recognition with the help of machine learning methods. It is very likely that LAWS will also be developed with machine-learning techniques. On the one hand, machine learning algorithms execute tasks such as face recognition more reliably than rule-based algorithms. On the other hand, they decrease the level of predictability in the machine’s functions because the underlying structure of the input-output relation cannot be known in all circumstances, which might lead to unanticipated emergent behavior (Trusilo, 2023). Thus, the machine might execute the task better than a rule-based expert system, but it does so with less predictability. Considering the environment in which LAWS will operate, this reduction in predictability might cause significant risks and dangers. However, this point received little attention when defining autonomous weapon systems. As discussed previously, Taddeo and Blanchard (2022b) analysis demonstrates that only two definitions, France and China, mention the use of machine learning in these systems. Machine learning provides adaptability to a novel situation in a real-world environment. So, the adaptive capacity when discussing the LAWS should be of significant interest. Merely following rigid rules has a twofold challenge when operating in real-world situations. First, real-world conditions are composed of many variables to represent in if/then rules, where the machine should only apply the *relevant* information to a given context.[8] As these systems lack the flexibility to adapt to new stimuli in warfare, LAWS with rule-based algorithms become more susceptible to errors in dynamic environments such as the battlefield. Therefore, the next subsection will tackle the predictability issue in rule-based and learner systems.

[8] This is conceptualized in the AI literature as the frame problem. For an analysis of the frame problem in the context of Arkin’s proposal of rule-based LAWS, see Klincewicz (2015).

3.4. (Un)predictability

Predictability means how much the actions of a machine can be anticipated. Holland Michel (2020) argues that there are two different aspects of predictability in autonomous systems: “technical” and “operational predictability” (2020, p. 5). Technical predictability is the feature of the machine. That is, it depends on the specific programming techniques employed in the systems. On the other hand, operational predictability refers to the complexity of the environment and situations in which the autonomous systems will operate.

Technical predictability relies on the system’s technological capabilities, such as the techniques for automating a specific task. For example, rule-based algorithms might be more predictable than machine learning algorithms adapting to environmental changes. However, the examples for rule-based LAWS provided previously are naively simplistic to address the predictability problem. Understandably, the examples are intended to illustrate the basic workings of such systems simplistically. In real-world cases, the complexity of tasks expected from LAWS would require extensive lines of code, potentially reaching millions. For comparison, F-35 fighter jets need “24 million lines of code” and “100 million lines of code for a modern luxury automobile” (Scharre, 2020, p. 166).

Another critical obstacle in deploying autonomous systems emerges when the system uses machine learning algorithms: explainability. Explanability poses a risk in the deployment of autonomous systems because of the difficulty in comprehending the underlying rationale behind the actions of these systems. In simple terms, explainability is the problem of not understanding why a system made a specific decision or acted in a particular manner. This problem is often referred to as the *black box problem* because the inner workings of these systems are not “transparent” (Diakopoulos, 2020, p. 197) but “opaque” (Burrell, 2016, p. 1). This does not mean that humans are entirely ignorant of the decisions. They may know the task assigned to the AI system. For instance, in face recognition, the programmers know that the system learns to recognize faces. Even though the program works well and recognizes faces reliably, this does not mean it is predictable. In black box cases, the programmers do not know how the system recognizes faces and what parameters or features the program uses. In that case, the process behind the outcome of face recognition is not explainable. Ultimately, this problem causes unpredictability because, without proper knowledge of the process behind its decisions, it is no longer possible to predict how the machine will behave in future scenarios. These considerations, however, refer to one type of unpredictability in autonomous systems that depends on the system's technological features.

Regardless of technological unpredictability, all types of autonomous weapon systems introduce some degree of "operational unpredictability" (Holland Michel, 2020, p. 5). Operational unpredictability is the problem of the operational environment. For example, autonomous weapon systems will have to operate in complex and dynamic environments with a plethora of inputs. Considering the wide range of inputs, such as friend-and-foe behaviors, other agents such as other military robots, geographical variations, weather conditions, and combinations thereof, LAWS are inherently unpredictable in an operational sense because it is not possible to envision or anticipate how the machine will interact with all these variables, and program the machine beforehand for all circumstances. To determine and predict all the future outcomes of an autonomous system "is not logically impossible, but it is unfeasible because the number of variables and their possible interactions is exorbitantly large, making this assessment intractable. (Taddeo & Blanchard, 2022a, p. 8). Thus, in an operational sense, unpredictability is a problem that cannot be easily overcome for both rule-based or machine learning systems because even if the explainable AI is achieved, meaning that it is possible to explain how the system made a specific decision or acted in a particular way, there remains operational unpredictability due to the difficulty in anticipating all potential situations that an autonomous system would encounter during its operations.

In conclusion, predictability in autonomous systems encompasses technical and operational senses, each posing specific difficulties. Technical predictability depends on the system's programming techniques and technological capabilities. Although rule-based algorithms may offer more predictability than machine learning methods, they both suffer from operational unpredictability. Operational unpredictability refers to the ability or lack thereof to navigate dynamic and unstructured environments with a multitude of variables. The critics of LAWS highlight the unpredictability aspect in relation to the inability to comply with the principles of *jus in bello*. They contend that the unpredictability of LAWS would constitute legal grounds to ban such systems. On the other hand, proponents contend that LAWS would reduce unethical behavior in warfare by minimizing the unpredictable behavior of human soldiers. Thus, regardless of which side one belongs to, technical and operational unpredictability must be addressed for a clearer understanding of the debate over the ethics and legality of LAWS.

CHAPTER IV

MORAL RESPONSIBILITY

Thus far, I have presented opposing views on the compliance of LAWS with the principles of *jus in bello*, more precisely, the principles of distinction and proportionality. Opponents argue that developing machines that comply with *jus in bello* is improbable because algorithms are inherently unpredictable. Thus, creating an algorithm to kill would likely increase casualties on the battlefield due to this unpredictability.

In contrast, proponents claim that once developed, these machines must be designed to ensure compliance with *jus in bello*. Furthermore, LAWS, they argue, have the potential to surpass humans' ethical conduct on the battlefield. This entails that these machines behave in ways that are more human than humans because, unlike human soldiers, they are not subject to physical, psychological, and cognitive limitations. As a result, there will be a reduction in casualties, war crimes, and military damage to non-combatant subjects and objects. This means it would be sufficient to develop machines that ethically outperform humans on the battlefield. Arkin himself points this out by claiming that "[i]t is not my belief that an autonomous unmanned system will be able to be perfectly ethical in the battlefield, but I am convinced that they can perform more ethically than human soldiers" (Arkin, 2009, pp. 30-31). Arkin acknowledges the argument that these machines may err and cause damages toward illegitimate targets, such as killing a civilian. Nevertheless, as long as they reduce human errors and crimes on the battlefield, LAWS will be preferable over the human soldiers who have displayed many historical examples of war crimes, inhumane treatment, and millions of civilian deaths.

Accepting such a view would even entail that it might even be an ethical obligation to employ such machines, as they would reduce the errors and crimes of war. However, there is a distinctive factor between errors of humans and errors of autonomous machines. When they commit crimes, humans are subject to the consequences of that crime, meaning that they bear the responsibility for the crimes. In the case of autonomous weapons, determining where the responsibility lies may be challenging and even lead to "responsibility gaps" (Matthias, 2004, p. 177). That is, it becomes difficult, if not impossible, to assign responsibility to any of the individuals involved in the morally loaded action carried out by a machine.

Responsibility assignment in LAWS' actions becomes important considering that responsibility is the fundamental presupposition for both *jus in bello* and IHL (Sparrow, 2007). As discussed previously, both *jus in bello* and IHL allow attacks only on legitimate targets (the principle of distinction), and noncombatants can be targeted if and only if the military advantage to be gained exceeds the harm and suffering imposed on illegitimate targets (the principle of proportionality). If responsibility assignments are abolished in war, then *jus in bello* principles would evaporate. This is because if no one is responsible, then there would no longer be anyone who should comply with them or be punished as a result of failing to comply with them. For this reason, war without responsibility would result in war crimes without criminals because there will inevitably be war crimes but no one to account for those crimes.

Responsibility assignment gets even more critical, considering that when machines fail, they might fail in ways such that we cannot estimate the risks associated with their failure. The abilities and advantages of these machines might result in catastrophic situations. For example, lack of physical, psychological, and cognitive limitations, such as resistance to fatigue and the ability to reach areas inaccessible to humans, are significant advantages, but how these abilities will contribute to the failures when these machines malfunction is a significant concern.

All these enhanced features may also amplify the consequences of their failures to a point where it is no longer comparable to human errors. Because the physical, psychological, and cognitive limitations of human soldiers will be eliminated in LAWS, they "could potentially kill for hours, with a death toll running into the hundreds, if not thousands" (Leveringhaus, 2016, p. 73). As a result, in this chapter, I will focus on the question of whether the use of LAWS would lead to a responsibility gap. While discussing the responsibility gap, a general overview of moral responsibility and conditions of responsibility assignments will be provided first. Then, I will proceed to the responsibility gap through possible loci of moral responsibility in the context of LAWS: the robot itself and the human agents, precisely the designers and commanders. After these considerations, I propose a possible solution to the responsibility gap within the context of LAWS.

Moral entanglement is the notion of moral responsibility in a vicarious sense. Vicarious responsibility is often practiced in everyday life but receives less attention when discussing responsibility gaps. The notion of vicarious responsibility acknowledges that in real-life situations, individuals can find themselves morally entangled in the actions and consequences of others, even if they do not have direct control or intent. This view recognizes a sense of moral responsibility that may not fit into traditional ethical frameworks, such as direct control over an action through causality, freedom, and epistemic conditions. However, vicarious responsibility accounts for a sense of responsibility through which individuals can be held morally responsible for the outcomes they do not have direct control over.

Ascribing responsibility in certain situations, particularly the ones involving advanced technologies, can be complex, but it does not necessarily follow from this complexity that ascribing responsibility is impossible. Thus, my argument aims to show that there are ways in which we can hold individuals morally responsible for the moral harm caused by LAWS. Therefore, the responsibility gap can be overcome.

4.1. Moral responsibility gaps

Technological advances introduce challenges to responsibility assignments due to the complexity of situations to which computers contribute significantly. Johnson & Powers (2005) argue that threefold factors shape the debate on techno-responsibility gaps. Techno-responsibility gaps "are ontologically,… conceptually,… [and] technologically complex" (Johnson & Powers, 2005, p. 99). Ontological complexity stems from the fact that multiple actors are involved in realizing morally loaded outcomes. For example, morally loaded action involving computer systems may include “modelers, coders, testers, documentation writers, system administrators, and users” (Johnson & Powers, 2005, p. 99). The increasing number of ‘hands’ involved in the morally loaded action is known as the “problem of many hands” (van de Poel, 2015, p. 50). The problem occurs when it becomes challenging to find individuals and their relations to a morally loaded outcome. Nissenbaum (1996) sees the problem of many hands as the characteristic of a computerized society, where many tasks in life are done with software systems that “are constructed out of segments or modules. Each module itself may be the work of a team of individuals,” (p. 39) and also computational systems are mainly developed “in institutional settings,… large corporations, government agencies and contractors” (p. 29). As a result, the number of individual actors involved increases the difficulty of assigning moral responsibility.

The second factor is technological complexity. Machines’ advanced technical capabilities pose challenges when assigning responsibility to human agents involved in morally significant outcomes (Matthias, 2004). As machines become more autonomous and capable of making complex decisions independently, it becomes blurry who should bear responsibility for their actions. If the machine operates based on rule-based algorithms, then the programmers and designers are morally responsible for the machine’s actions. As discussed in the previous chapters, in cases where the machine incorporates learning and adaptive algorithms, their action can be less predictable and more difficult to attribute to their programmers.

Conceptual complexity is the third and final factor shaping the responsibility gaps. The diversity of perspectives on responsibility paves the way for the different conceptualization of responsibility gaps. In the present thesis, the moral responsibility gap is the central concern (not, for example, legal responsibility gaps), and moral responsibility is particularly understood as blameworthiness. In this sense, being morally responsible for an action means that the agent is also blameworthy and should account for the action. The reason I take responsibility as blameworthiness as the viewpoint for my analysis is the fact that the moral debate surrounding the responsibility gap focuses on the culpable aspect of moral responsibility (Santoni de Sio & Meccani, 2021; Königs, 2022). Before concentrating on moral responsibility gaps, however, we need to have an understanding of moral responsibility in general.

4.2. What is moral responsibility?

The metaphysical discussion on free will and determinism has been one of the most central issues in the general debate on moral responsibility (Talbert, 2022). In its essence, the debate on free will focuses on the view that if determinism is true, we do not have free will. In a moral context, this would mean that we cannot hold anyone responsible because they could not act otherwise. After all, the law of nature determines all actions. If we cannot control our actions or refrain from specific actions, then we cannot be responsible for that action or omission. Intuitively, free will seems to be the necessary condition to be held morally responsible. However, the thesis of determinism claims that an agent acts in a way not because of free will but because of the forces pre-determined by the law of nature. This holds that not only forces by other agents but also ones such as neurobiological determinations lead to a specific course of action. Accepting such a view could leave the moral responsibility assignments unjustifiable because the law of nature determines the course of action one would take regardless of whether the agent intends to take those actions. While the free will/determinism discussion is an important endeavor, the free will debate is outside the scope of this present thesis. In fact, Strawson thinks that our interpersonal relationship concerning moral responsibility is not necessarily contingent upon "a general theoretical conviction [about determinism]" (2008, p. 12). This means that even if the advances in sciences such as neuroscience show that physical or neurobiological causes pre-determine all our behaviors, our interpersonal practices, such as holding people morally responsible, would not necessarily change.

In the Strawsonian view, our interpersonal relations consist of two kinds of attitudes: the "objective" and "reactive attitudes" (Snowdon & Gomes, 2023, section 8, para. 2). Objective attitudes are aligned with the thesis of determinism, where we are inclined to have such attitudes towards individuals whom we perceive as "incapacitated in some or all respects for ordinary inter-personal relationships" (Strawson, 2008, p. 13). For example, someone suffering from a psychiatric condition is a target of an objective attitude and is treated as "an object of social policy" (Strawson, 2008, p. 9). Psychiatrically decapitated agents are perceived as persons in need of treatment, so they are subject to deterministic mechanisms but not appropriate targets of moral responsibility, praise, or blame. On the other hand, our reactive attitudes are preserved for individuals perceived as appropriate targets of some reactions such as resentment, indignation, blame, praise, forgiveness, and so on. (Strawson, 2008). According to this view, someone is a morally responsible agent if and only if they are an appropriate candidate for certain types of reactive attitudes, such as blame or praise. Thus, building upon the Strawsonian view, I will also confine my analysis of moral responsibility and determinism to the *assumption* that moral responsibility and holding people responsible are so embedded in our ordinary life that it is hardly conceivable that the truth of determinism would alter their centrality in everyday experience.

This assumption, however, does not answer the question of how reactive attitudes are appropriately assigned to others' behaviors. What reactive attitudes offer is that our moral responsibility practices do not depend on the metaphysical considerations of free will and determinism, and moral responsibility assignments are integral to ordinary life. However, under what conditions we have certain reactive attitudes towards someone remains unclear. This question is a separate but highly related issue to the compatibilism of free will and determinism because, since Aristotle, the common views on the conditions of moral responsibility depend on the free and voluntary actions of the agents.

4.2.1. Conditions of moral responsibility

In Nicomachean Ethics, Aristotle claims that we can blame or praise someone for their "voluntary actions, i.e., actions done not by force, and with knowledge of the circumstances" (1109b/30-35)[9]. This account thus proposes that a morally loaded action, that is, an action that can be characterized as either praiseworthy or blameworthy, requires an agent to do an action voluntarily, which is to act freely and knowingly. Aristotle's views on blame and praise have been mostly adapted and minimally modified for the discussion of moral responsibility, and generally, at least three conditions are accepted as a consensus for an agent to be counted as a morally responsible agent: the epistemic, causal, and freedom condition (Noorman, 2023). The conditions define how an agent relates to an action, and if the conditions are met, the agent is a morally responsible agent; therefore, s/he is an apt candidate for blame. Firstly, the epistemic condition for assigning responsibility requires that the agent is

[9] David Ross(trans.), Oxford University Press, p. 38.

aware of the situation in a suitable manner. It would not be appropriate to hold someone responsible for an action if they did not know or could not have known the consequences of their actions. For example, suppose Bill has a bomb mechanism embedded in his phone. Unbeknownst to him, when someone calls him, the bomb activates and leads to Bill's death. Bill's friend, Alex, calls him, causing the bomb to explode and killing Bill. In this case, can we hold Alex responsible for Bill's killing? It is important to note that Alex did not know that making a phone call would trigger the bomb. Since she was unaware of the situation and lacked the requisite information, holding Alex responsible and blaming her for Bill's death would be morally unfair.

Secondly, the causal condition for assigning responsibility requires the agent to be in a causal connection with the morally wrongful outcome. Causality is essential because it would not make sense to hold someone responsible for an action to which they have no causal relation. Causal responsibility, however, does not mean that the agent is morally responsible. An earthquake, for example, might be causally responsible for the deaths of many people, but it would not make sense to hold the earthquake *morally* responsible and suitable to be blamed.[10]

Alternatively, imagine that your friend accidentally spills hot water on you. This incident might evoke certain reactive attitudes towards your friend. However, your reactive attitudes in this situation differ from those you would hold towards someone who consciously intends to pour hot water on you. In both cases, there is causal responsibility for the outcome of 'you getting burned,' but the distinction in the two cases lies in the fact that your friend did not have control over the action; in the latter case, the person intended to harm you with hot water.

This latter point brings us to the last condition: the freedom condition. The freedom condition includes the agent's intents, desires, and wants. If the agent is coerced by someone or compelled by forces other than their own will to do X (e.g., action or omission), then we cannot hold the agent responsible for X. While there are contested views on all three conditions, the freedom condition is arguably the most debated one

[10] Natural disasters might be blamed in cultures where the spiritual system of that culture allows one to attribute moral agency to natural disasters. Suitable candidates for reactive attitudes, therefore, are greatly dependent on the culture one lives in.

because of the challenges from the thesis of determinism. Also, what the intents, wills, and desires mean is not philosophically clear.[11]

As discussed previously, the views on moral responsibility in the present work depend on the Strawsonian assumption that moral responsibility and the freedom condition, particularly, are integral parts of interpersonal relations in ordinary life. This perspective avoids discussing the extensive debate surrounding determinism, free will, intentions, desires, and related topics. It is worth noting that not only the freedom condition but also both the epistemic and causality conditions have rich and diverse philosophical debates associated with them. However, the primary focus of this thesis is more practical in nature, emphasizing real-world implications of moral responsibility within the context of LAWS rather than delving into the philosophical considerations of the above-referred topics. As a result, the thesis will not challenge the three commonly held conditions of moral responsibility. Thus, satisfying these conditions –meaning that the agent causes a morally wrongful outcome freely, knowingly, and causally– will suffice to claim that the agent becomes morally responsible and blameworthy for the resulting consequences.

4.3. Moral responsibility and LAWS

In his seminal paper *Killer Robots*, Sparrow (2007) argues that it is impossible to hold anyone morally responsible for the actions of LAWS, thereby leading to a responsibility gap. He sees that the responsibility gap is the most important reason that the deployment of LAWS would be unethical because "it is fundamental condition of fighting a just war that someone may be held responsible for the deaths of enemies.… In particular, someone must be able to be held responsible for civilian deaths" (Sparrow, 2007, p. 67).

As previously discussed, responsibility gaps emerge in situations where a morally significant outcome occurs. Yet, it becomes difficult, or even impossible, to identify an individual who satisfies the conditions of moral responsibility and can be justly held responsible for that outcome. Thus, for a responsibility gap to occur in the case of LAWS, there should be no individual who either knew or could have reasonably

[11] For a philosophical view on intentions and responsibility, see, for example, Mele & Sverdlick, 1996.

known the consequences of the actions of LAWS (epistemic condition). The same agent should lack control over the actions of LAWS in the sense that they would not intend to cause the consequences or could not have prevented them from occurring (causal and freedom conditions).

Within the context of LAWS, the actors behind the actions of LAWS can be understood and allocated across three layers. At the first layer, we have the machine itself, as it directly engages in carrying out the actions. The second layer is the users or deployers of the machine. This could be a human commander who made a decision with certain intentions to employ the machine. Finally, the third layer is the designers of the machine.[12] The designers' intentions are allocated to the design, development, and programming of the machine.

Identifying these three layers as potential loci of moral responsibility for the actions of LAWS, I will now address Sparrow's argument asserting that none of the agents within these layers satisfy the conditions of moral responsibility.

4.3.1. The robot

At first glance, it may appear that it is nonsensical to ascribe moral responsibility to the robot itself for its actions. However, it is useful to understand the reasons why discussing the moral responsibility of machines in their actions is unfeasible and that we cannot assign responsibility to them. Since conditions of responsibility are commonly perceived as capacities exclusive to humans with consciousness, the notion of holding machines responsible for their actions appears counterintuitive. Indeed, Levy (2014) argues that consciousness is the most important prerequisite for moral responsibility. Sparrow(2007), in a similar vein, poses consciousness as a necessary prerequisite of moral responsibility, albeit in a more specific manner, asserting that moral responsibility requires the capacity to experience emotions such as guilt and suffering. His argument depends on the claim that,

> X is considered morally responsible if and only if three conditions are met: (i) X is an appropriate candidate for blame; (ii) it is conceivable to impose

[12] I use the term "designers" to refer to the agents involved in the design, programming, and development stages, rather than using the terms "designers," "programmers," and "developers" separately.

> punishment on X; (iii) it is conceivable for X to suffer as a result of punishment.[13]

This means that for robots to be considered morally responsible, they must fulfill the three conditions. First, in (i), the "appropriate candidate for blame" is subject to diverse interpretations and conditions set for moral responsibility, and consciousness might be conceived as the necessary condition to be an appropriate candidate for blame. Second, in (ii), Sparrow subscribes to a retributivist perspective on punishment – the notion that the wrongdoers deserve punishment as a consequence of their actions. Finally, in (iii), for Sparrow, the most plausible account of punishment is that punishment "requires that those who are punished, or contemplate punishment, should *suffer* [emphasis added] as a result" (Sparrow, 2007, p. 72).

Even if we were to set aside the first two conditions, Sparrow highlights the importance of consciousness in the third condition, specifically that punishment necessitates the capacity to suffer. As of now, it is not plausible to expect these conditions to be met by robots, given the absence of subjective experience of suffering in robots. Nevertheless, some argue that advanced learning capabilities and autonomous capabilities will make machines appropriate candidates for moral responsibility. For example, Hellström (2013, p. 105) claims that,

> [A]dvanced learning capability will not only make it harder to blame developers and users of robots, but will also make it more reasonable to assign responsibility to the robots. If a robot learns and changes behavior as a result of praise and blame it receives, it may actually make sense to ''punish'' the robot.[14]

An important distinction between Sparrow and Hellström lies in their conceptualization of punishment. Hellström's proposal does not depend on the belief that the wrongdoer should be punished and suffer; instead, it centers on the idea that the robot modifying its behavior to prevent the recurrence of the same wrongful actions constitutes a justification for punishment. Hellström thinks of punishment as a means

[13] This is not a general description of moral responsibility; rather it is the reformulation of Sparrow's argument. This section is an analysis of responsibility gap in LAWS as proposed by Sparrow.

[14] Here Hellström refers to reinforcement learning - one of the methods in machine learning. In reinforcement learning, machine takes several action and receives a reward that indicates how good or bad the action was. As a result, the machine tries to maximize the reward by taking the best actions. (For more detailed account of reinforcement learning, see, Alpaydın, 2016).

to an end. For this view, punishment is a deterring means for achieving the end that the same wrongs do not occur in the future. Here, punishment is conceived as a preventive effect on people because it is assumed that people avoid suffering; therefore, they deter from misdeeds. If we can achieve the end that an agent is deterred from a misdeed, then other methods would be as plausible as imposing punishment, such as an update in software or hardware of a robot.

Although very much debated and criticized by the proponents of alternate methods, such as rehabilitative theories, retributivist punishment is both common-sensical and supported by many moral philosophers (Danaher, 2016). Moreover, Danaher, drawing on both ethnographic and psychological evidence, claims that "humans are innate retributivists" (2016, p. 299). The moral plausibility of punishment, conceived as a robot altering its behavior in response to praise and blame, may appear somewhat unintelligible in the context of LAWS because this view would amount to informing the victims of war crimes that "the responsible robot has been punished". How plausible this punishment is in the context of serious war crimes is questionable. The burden of proof falls on the proponents of this view to demonstrate that such a form of punishment also meets the needs and expectations of the victims of wrongful acts.

Furthermore, suffering is not only essential to punishment, but for robots to have moral status, they would have to possess "psychological and social properties, such as capacity for rational thought, pleasure, pain, and social relationships" (Schwitzgebel & Garza, 2015, p. 101). Therefore, conceiving punishment as correcting one's behavior in accordance with reward and punishment and that robots satisfy this condition would not be sufficient to claim that robots would also have moral status. As of now, machines do not possess psychological and social capacities, such as suffering, and it is hardly possible to envision how a robot could be subjected to traditional forms of punishment.

Let us set aside such questions and rather assume the prospect that robots will be appropriate candidates for moral blame and that they will even have the capacity to experience emotions or convincingly simulate emotions, especially suffering.

Affective computing is a specialized field that investigates the potential of replicating human affects in machines (Picard, 2000). In the context of LAWS, for example, a machine could simulate the experience of pain as a consequence of punishment. However, even if affective machines were capable of simulating emotions to a high degree, it remains an open question whether this would lead to them being regarded as moral agents and whether it is desirable to develop affective machines within the context of LAWS.

There are three problems pertaining to the development of affective LAWS. Among these, two of them appear to be in contradiction with the initial motivations that drive the development of LAWS. The third is a broader concern regarding affective machines in general, which extends beyond the specific context of LAWS.

The first problem is that if robots have advanced to the point where they can experience emotions and participate in social interactions as part of human moral practices, their losses in warfare will evoke similar emotional responses as those experienced when human soldiers are lost on the battlefield. This is so because LAWS "would have become our soldiers, and we should be as morally concerned when our machines are destroyed -indeed killed- as we are when human soldiers die in war" (Sparrow, 2007, p. 73). This contradicts the initial motivation to develop LAWS in the first place, which is to reduce the loss of human soldiers. If we were to experience the same empathetic emotional responses to the "killing" or destruction of LAWS as we do for human casualties, then the consequentialist advantage of reducing human soldier casualties is diminished. This is because the emotional responses such as sadness, pain, anger, and desire for revenge would still remain.

The extent of our emotional responses to the "death" of a robot compared to the ones we might have for the loss of humans is unclear. However, this question should be addressed if one is to argue that LAWS can become suitable candidates for blame and that they can be developed to be held morally responsible and "suffer" as a result of their actions.

Consequently, it raises the question of whether affective LAWS should be developed in such a way that they can be apt candidates for moral responsibility and held responsible for their actions through punishment, but not to the extent that we develop strong emotional connections or states toward them. This balance between responsibility and emotional attachment is a critical ethical consideration in the development of advanced LAWS. The second problem, which is relevant to the aforementioned problem, is about the question of whether equipping LAWS with emotions like pain, guilt, anger, etc., would impair their judgment. It has been argued that one of the most significant drawbacks of human soldiers is their susceptibility to emotions, which can lead to poor decision-making and result in atrocities and war crimes. In contrast, LAWS that are devoid of such emotional states might be considered more human than humans because they would not be subject to emotions that "cloud their judgment" (Arkin, 2009, p. 29). Thus, the paradox remains: LAWS with emotions could be considered morally responsible, but introducing emotions to them might compromise their effectiveness and ethical behavior on the battlefield due to the influence of these emotions.

Finally, the third problem is a rather broader ethical question about the morality of creating machines with emotions, not primarily in relation to the harm they might inflict on humans as raised in the previous problem, but rather in terms of the pain and suffering these machines themselves could experience. Wallach and Allen (2009), for example, express this concern:

> If robots might one day be capable of experiencing pain and other affective states, a question that arises is whether it will be moral to build such systems —not because of how they might harm humans, but because of the pain these artificial systems will themselves experience. In other words, can the building of a robot with a somatic architecture capable of feeling intense pain be morally justified and should it be prohibited? (p. 209)

To conclude, the development of affective LAWS for the purpose that they might be part of our moral responsibility practices and, as a result, can be blamed and punished introduces complex problems. Within the context of LAWS, affective machines would lead to contradictions between the original rationale to build LAWS and the consequences of equipping them with emotions.

In my analysis, two contradictions have been pointed out. The first is about the claim that LAWS would reduce the loss of human soldiers and, therefore, reduce the emotional burden of losing humans on the battlefield. However, the prospect that robots could be held morally responsible and punished would also entail that there is a degree to which humans will develop other emotional responses towards robots. The dilemma in this problem stems from the claim that robots would have sufficient emotions to be held responsible and yet not to such an extent that it triggers other emotional responses in humans, including feelings of sadness, anger, and pain as a result of the "death" or "injury" of these robots.

The second is that, unlike social robots such as caregiving robots, integrating human-like emotions into LAWS can lead to even greater risks, including the potential for acts of vengeance, fear, cowardice, etc. The dangers associated with LAWS exhibiting emotions in the context of warfare outweigh any potential benefits of developing them with emotions for the purpose of punishment. Thus, while affective computing can enable certain autonomous machines to simulate emotions, developing LAWS with emotional intelligence does not offer a viable solution to the issue of holding them responsible. The third is a broader ethical dilemma surrounding the question of whether it is morally justifiable to build robots with the capacity to feel pain and other emotional states.

As a result, in order for a robot to be held morally responsible, it should either experience emotions or express emotions as if it experiences them. However, this expression should be "in ways that will establish the moral reality of these states" (Sparrow, 2007, p. 72). In other words, robots must possess a level of emotional capacity for us to impose punishment on them as a result of their blameworthy act. If machines lack this emotional capacity, then it is not intelligible to hold them morally responsible. Moreover, I have also argued that it is not ethically desirable to build LAWS with affective capacity because such machines are contradictory to the initial rationale to deploy them on the battlefield – the view that LAWS are more advantageous than humans because they are not subject to emotional states which blur the judgment of human soldiers in the battlefield.

4.3.2. Human agents: The commander and the designer

After the direct causal relation between the robot and its actions. The second and third layer consists of deployers and developers of LAWS. Regarding their roles in the consequences of the robot's action, deployers and developers in the actions of LAWS may have control over LAWS to some extent. Deployers of LAWS could be commanding officers who order the robot to execute tasks in certain geographical locations and for certain periods of time. Before deploying the machine, they have the decision-making authority to limit the operational space and time of the machine. Holding deployers responsible would be fair if they choose to deploy the machine in a geographical area that falls outside the machine's designed scope, as they willingly assume the associated risks of deploying a machine outside of the design parameters.

The deployers can provide general instructions and objectives for the machine, but each individual action the machine will take is unpredictable and beyond the commanding officers' control. Consequently, holding commanders responsible for the machine's actions would be unjust. Sparrow claims,

> The autonomy of the machine implies that its orders do not determine (although they obviously influence) its actions. The use of autonomous weapons therefore involves a risk that military personnel will be held responsible for the actions of machines whose decisions they did not control. The more autonomous the systems are, the larger this risk looms. At some point, then, it will no longer be fair to hold the Commanding Officer responsible for the actions of the machine. If the machines are really choosing their own targets then we cannot hold the Commanding Officer responsible for the deaths that ensue. (2007, p. 71)

In a similar vein, the third layer, developers of LAWS, would not be held morally responsible because "[t]he connection between the programmers/designers and the results of the system, which would ground the attribution of responsibility, is broken by the autonomy of the system" (Sparrow, 2007, p. 70). In other words, the robot with autonomous capabilities will take an action that is neither intended nor reasonably foreseeable by its developers.

As discussed in the previous chapters, programming a machine for complex tasks introduces uncertainty, as the machine may exhibit behaviors that were neither anticipated nor intended by its programmers. The uncertainty is even more at stake when machine learning algorithms are employed, as the machine learns from and adapts to its environment. Even if a machine is programmed with a set of general rules to abide by, which the proponents of LAWS propose, it is still hardly possible to foresee how these rules will be applied in particular situations. While the machine generally operates within the rules, there will be instances where the application of these rules leads to outcomes that the programmer cannot foresee. Therefore, it would be unfair to hold programmers responsible for actions that they did not have control over.

In considering the responsibility for the actions of LAWS, developers, along with commanding officers, can only fulfill the causality condition. While these individuals may be causally related to the actions of the LAWS, they do not satisfy the other two conditions, freedom and epistemic conditions. They may have neither intentions for the morally harmful outcome nor complete knowledge of the consequences resulting from the robot's actions.

Ultimately, for Sparrow(2007), these considerations give rise to a responsibility gap. On the one hand, we cannot hold the machine itself responsible for its actions as it lacks the necessary conditions for moral agency. On the other hand, humans involved in the operations of LAWS may no longer fulfill key conditions of moral responsibility, such as having knowledge of the consequences and having intents behind the actions of the machine. Therefore, this leads designers and deployers to defend themselves on the moral grounds that they do not have control over the harmful actions of LAWS.

4.4. Bridging the gap

After analyzing the responsibility gap within the context of LAWS, I will now turn to a specific solution for addressing this gap. In the subsequent part of the chapter, I will present a potential solution to bridge the responsibility gap in LAWS by exploring an alternative interpretation of moral responsibility, known as vicarious responsibility. Vicarious responsibility arises in situations where one agent bears responsibility for another's actions, even if that agent did not have direct control over the actions of another. In this section, I will demonstrate how designers of LAWS, contrary to Sparrow, can be held morally responsible for morally harmful outcomes of LAWS.

4.4.1 Direct control and responsibility

Previously, it has been proposed that an agent is morally responsible on the condition that that agent fulfills causal, freedom, and epistemic conditions. Once the agent satisfies all these conditions, then the agent is morally responsible. The responsibility gap argument, thus, claims that nobody satisfies these three conditions in the harmful conduct of LAWS. The argument against the responsibility gap would proceed as follows:

(i) If there is an agent retaining control over the actions of LAWS, then there is no responsibility gap.

(ii) An agent retains control over the actions of LAWS.

(iii) Therefore, there is no responsibility gap.

In order to reject the responsibility gap, there have been several attempts to show that (ii) an agent has control over the actions of LAWS. The common feature of arguments for control is that they aim to show that control does not necessarily require "direct controlling." (Santoni de Sio & van den Hoven, 2018, p. 10).

The absence of control over the actions of LAWS is assumed particularly as the absence of *direct control*. The physical distance between the outcome and the agents seems to indicate that agents lack control over the outcomes. In this context, I have also referred to layers in the actions of LAWS. I explained that there is a direct causal relationship between the machine and the consequences of its actions, while operators/commanders are in the second layer, and designers are in the third layer. However, the distance in time and space does not mean that the agent loses control over the outcomes, "as control in a morally relevant sense allows for technological mediation and separation of the human agent and the relevant moral effects of the acts that he [she] is involved in" (Santoni de Sio & van den Hoven, 2018, p. 10).

4.4.2. Vicarious responsibility

Moral responsibility generally involves a direct relationship between the agent and the action, meaning that there is a minimal distance in terms of both time and space between the agent and the action. However, in many cases, there can be no direct link between the agent and the action. Let us suppose that you are invited to your friend's house, and you take your dog with you. While visiting your friend, who has a sketchbook full of drawings she has created over the years, your dog ends up damaging the notebook by biting it. Who would be responsible for the damage in this scenario: the dog itself, you, or your friend?

In situations like this, even though you may not have direct control over your dog's actions in the way you would over your own, there is still a sense of moral responsibility on your part. You are supposed to respond to the situation distinctively and differently from another person who is just a bystander. For instance, you would be expected to offer apologies to your friend, try to find ways to compensate for the damage, take precautions to prevent your dog from causing similar incidents in the future, and so on. Put shortly, you are expected to take responsibility for the actions of your dog. In a similar vein, if you fail to take responsibility, such as not responding in the ways exemplified above, there is a moral sense that you react to the situation in a morally inappropriate way.

This understanding of moral responsibility is vicarious responsibility, where an agent bears responsibility for the actions or behaviors of another entity (Mellor, 2021; Goetze, 2021; Glavanicova & Pascucci, 2022). This other entity can be another human, a non-human animal, a collective, or a robot. Vicarious responsibility emerges between two entities because of a special type of relation one has with another entity, which Goetze calls "moral entanglement" (2021, p. 220). So far, there is still an obscurity of that special moral relation between two agents that allows one to take responsibility for another's actions. For example, let us consider the case of the designers, as they are seen as the most appropriate candidates for moral responsibility in the outcomes of autonomous systems (Goetze, 2022; Taylor, 2021; Gotterbarn, 2001).

The designers influence the actions of an autonomous system. The designers' intentions, in some abstruse sense, are present in the LAWS. Goetze, for instance, claims that the intentions of the designers become apparent in their control over "when the training [of machine learning system] has been successfully completed,… choosing the training dataset, creating the reward function, tuning the hyperparameters, and so on" (2022, p. 9). In addition to the software, designers' choices of hardware, such as the type of the munition (bomb, bullet, etc.), contribute to the moral entanglement of the designer's agency with the machine and its actions. In addition to these choices, designers of autonomous systems will often be required to address and take actions to rectify the harmful behaviour by, for instance, updating the software or hardware of the system (Goetze, 2022).

Thus, two aspects of moral entanglement appear. First, the intentional decisions made by the designer regarding the software and hardware components of the system contribute to the moral entanglement of the designers with the machine. Second, their expertise makes the designers appropriate candidates for correcting the harm the machine causes, and as a result, makes them somewhat related to the LAWS' harm. Thus far, these two aspects give a sense of moral entanglement between the designer and the LAWS.

However, it seems plausible to claim that this analysis is far from being clear. Goetze also accepts this obscurity of moral entanglement by claiming that in real-world scenarios, "what we are personally responsible for are often genuinely unclear" (2022, p. 8). Thus, there remains uncertainty and obscurity in the explanation of moral entanglement that gives way to vicarious responsibility.

In what follows, however, I aim to provide a clarification for the claim that designers of LAWS are vicariously responsible for the machine's actions. For this purpose, I will draw upon Glavanicova & Pascucci's (2022) definition of vicarious responsibility.[15] According to this definition, agent A is vicariously responsible for Q if and only if

– (i) Q is a morally harmful outcome in a set X;

– (ii) an agent B causally contributes to bringing about Q;

– (iii) A is voluntarily involved in a relation R with B;

– (iv) the set X falls within the scope of the relation R.

To better grasp this definition, let us consider a hypothetical case of LAWS killing a surrendering soldier (an illegitimate target):

-(i) 'the surrendering soldier is killed' is a morally harmful outcome in a set 'some target is killed or no target is killed'[16]

[15] For our purposes, I have modified Glavanicova & Pascucci's account. The original formulation of their analysis is as follows: "A normative party A is vicariously responsible for a proposition P if and only if (i) P instantiates a prohibited proposition in a set X; (ii) an entity B causally participated in bringing about P; (iii) A is voluntarily involved in a relation R with B; (iv) the set X falls within the scope of the relation R" (2022, p. 17).

[16] Here I used Himmelreich's distinction between particular and general outcome. He defines "Outcome A: this particular target is bombed; Outcome B: some target is bombed or no target is bombed" (2019, p. 738). Outcome A is a strict subset of outcome B. Himmelreich claims that commanders would be responsible for the outcome B, where outcome A is not intended in and of itself but intended as a risk. Taking risk entails responsibility. Therefore, commander would have to justify why they took a risk by giving an order. Although Himmelreich does not focus

-(ii) LAWS causally contributes to bringing about ‘the surrendering soldier is killed’

-(iii) The designer is voluntarily involved in a relation ‘design’ with LAWS

-(iv) the set ‘some target is killed or no target is killed’ falls within the scope of the relation ‘design’

Therefore, the designer is vicariously responsible for ‘the surrendering soldier is killed.’

In this instance, (i) means that the surrendering soldier’s death is within the potential outcomes that LAWS can cause, i.e., some target is killed or no target is killed. (ii) refers to the causal contribution of LAWS to the event. LAWS kills the soldier, so it is in a causal relation with the soldier’s death. The (iii) is particularly important because, as discussed above, the designer’s intentions and choices over the software and hardware of the system during the design phase influence, though not entirely determine, how LAWS operate in warfare. This also entails the designer’s voluntary involvement in the corporation that manufactures LAWS. For (iv), the morally relevant relation is the ‘designing’ relation between the designer and the LAWS. The general purpose of manufacturing LAWS is to engage targets, and this entails that LAWS will engage some targets during their operation. Note that the designer neither had direct control over the particular outcome nor was s/he aware that LAWS would engage that particular target. However, there is still control of the designer in the morally relevant sense over that particular outcome.

Consequently, these considerations demonstrate the ways in which the designers can be held morally responsible for LAWS’ actions. This refutes the responsibility gap argument by proving the premise that some agents retain control, albeit indirectly, over the actions of LAWS. Therefore, the responsibility gap can be overcome.

However, the above clarification does not aim to argue that, in all circumstances, the designers would be responsible. There may be particular cases where, for example, moral harm is caused by many other factors that are outside the scope of the relation

on vicarious responsibility. The commander’s case is also applicable to the definition. ‘some target is bombed or no target is bombed’ falls within the scope of ‘ordering to deploy LAWS.’ Thus, commanders would be vicariously responsible.

design. In order to clarify what the scope of the design relation entails, some further remarks must be made before closing the present chapter.

4.4.3. Scope of vicarious responsibility

The above-proposed solution might not be applicable in all circumstances that a LAWS engages in a morally harmful action. For instance, if a commander knowingly deploys LAWS in geographical locations where they are not designed for, and the system causes a harmful outcome. In this case, a responsibility gap would not emerge because the commander, by deploying the machine in a situation that is outside of the systems' design parameters, would be morally responsible for the morally harmful outcome.

Another objection would be to clearly define the scope of design relation, such as written agreements, so as to limit the vicarious responsibility of designers. Although it is not a fully developed counter-argument, Sparrow(2007) has raised this objection to holding designers responsible. According to him, it would be even more difficult to hold designers responsible if they "acknowledged the limitations of the system" (2007, p. 69). If designers explicitly acknowledge the limitations of their design and this acknowledgment is clearly defined in written agreements, it could impact the scope of the relation 'design'. This means that if the scope of the design relation is determined by written agreements with the buyers, then this could exempt designers from responsibility for morally harmful outcomes. However, limiting the scope of design-work to acknowledging the limitations of the system could only exempt designers from legal responsibility. For instance, limiting the scope of the design relation by agreements could protect designers from certain legal liabilities, such as paying compensation to the victims. However, the acknowledgment of the limits of their design does not necessarily absolve designers' moral responsibility. Thus, I think that this objection would only be the case for legal responsibility, but accepting the limitations in their design would not exempt designers from moral responsibility. Moreover, this type of exemption from moral responsibility would also require moral justification of playing a "moral gambit" on the lives of the innocent (Taddeo & Blanchard, 2022a, p. 17). In other words, the agreement would mean that the designers willingly take the moral gambit on the lives of others and violations of ethical principles in warfare. The scope of design relation, even if designers acknowledge the limitations, would still entail the moral harm resulting from LAWS' actions. Thus, their vicarious responsibility would still remain.

Another more challenging example would be the case of group responsibility. At the beginning of this chapter, I mentioned that three complexities are contributing to the responsibility gaps in technologically mediated outcomes: conceptual, technological, and ontological complexities. The ontological complexity constitutes a particular challenge for holding an individual responsible because there are many agents contributing to the outcomes mediated by technological systems.

One could claim that the vicarious responsibility solution to the responsibility gap remains problematic for real-world scenarios concerning military decisions, where the factors contributing to moral harm are often caused by the decisions made by multiple layers distributed within and beyond military organizations (Schulzke, 2013). Taylor (2020), for instance, suggests that "a number of distinct groups might be identified as potential loci of responsibility: the government, the military, and the developers of LAWS" (p. 327). Given this distributed nature of technologically mediated outcomes in the military context, it would be unfair to expect an individual to cover "the full gravity of the moral harm done" (de Jong, 2020, p. 732). This objection is even stronger when considering the fact that technological artifacts are often developed by multiple individuals and organizations, where it is not plausible to pinpoint one individual responsible for the whole design process. Because of this ontological complexity, it could be argued that responsibility should either lie on the collective as a whole or should be shared by the individuals in the design, development, and deployment stages of LAWS (Floridi, 2016; Taddeo & Blanchard, 2022a).

However, neither of these positions refutes the vicarious responsibility of designers. My argument is not that the designers are the *only* candidates for moral responsibility. Instead, I have aimed to show in what sense designers can be held morally responsible. Designers' vicarious responsibility can be incorporated into the group's moral responsibility for LAWS actions. The group agency in the context of LAWS would consist of individuals involved in LAWS' design, development, and deployment stages. Since the morally harmful outcome of LAWS occurs due to the decisions made in the multiple layers of these stages, as the argument goes, the group is perceived as the agent of the outcome and morally responsible for that outcome. Designers are members of the group that is the morally relevant agent for the outcome and are necessary members of that group to be the agent of morally harmful consequences. More precisely, my proposal explains why designers must be a necessary member of the group agent that is morally responsible for the outcome. Therefore, the vicarious responsibility of designers would align with the view that responsibility should either lie in the collective as a whole or be distributed to individual members of that collective.

CHAPTER IV

CONCLUSION

In this thesis, I investigated the moral problems of lethal autonomous weapon systems (LAWS). The main concern of the thesis is the question of whether LAWS lead to responsibility gaps. I have argued that it is possible to hold designers of LAWS responsible, albeit in a vicarious sense, for the conduct of these systems. Thus, the responsibility gap problem can be resolved.

In Chapter 2, I analyze different definitions and frameworks used in the literature to define LAWS. The chapter shows that there is confusion over the descriptions of LAWS. This problem occurs because there is limited knowledge about the precise nature of LAWS. Because of this, the debate is primarily speculative; it relies on predictions about future technology. However, to better understand LAWS, I have first analyzed the often-cited definitions of the US Department of Defense and the International Committee of the Red Cross. Both of these definitions highlight an aspect of autonomy in weapon systems: functioning without human intervention.

After discussing these definitions, I have analyzed autonomy in machines. Autonomy in machines refers to the capacity of a machine to operate independently of human intervention in its operation. However, this is a limited view of autonomy in machines because this understanding would include machines executing relatively simple tasks. For example, household appliances such as washing machines also work independently of human intervention after starting the machine. Thus, further parameters have been discussed: the complexity parameter and the type of functions automated.

The former refers to the three-dimensional complexity in autonomous systems, and the latter is about the question of which functions of a system, when automated, make the system as a whole autonomous. The threefold complexity that makes autonomous machines different from the ones that execute relatively simple tasks consists of the complexity of the machine, the complexity of the task assigned to the machine, and the complexity of the machine's environment. The interrelation between these three complexities contributes to the autonomy of machines. The first refers to the inherent complexity of the machine, more particularly the advanced software and hardware features. The second is the complexity of the task assigned to the machine, for instance, the difference between the tasks stopping the washing cycle in washing machines and automated braking when a pedestrian is detected in self-driving cars. The third factor, related to the aforementioned complexities, refers to the environmental difficulty, where the difference between structured environments and unstructured environments becomes necessary for assessing the autonomy of machines. Along with the machine's complexity, task complexity, and environmental difficulty constitute the complexity parameter in machine autonomy assessments.

After the complexity parameter in autonomy in machines, I have discussed the third parameter: the type of decisions or functions automated. This parameter refers to a shift in the perception of autonomy in machines. It suggests that instead of viewing machines as autonomous in their entirety, it emphasizes the importance of focusing on the specific decisions or functions when automated make a machine autonomous. For instance, significant attention in autonomy in weapon systems debate has been given to the automation of killing function while functions such as target recognition (i.e., identifying and prioritizing targets but not actual killing) have been automated for a long time. Thus, according to this parameter, autonomy is not a blanket characteristic but rather a quality that emerges when certain functions of a machine's operation are automated.

After analyzing autonomy in machines, I turn to the discussion on autonomy in weapon systems. In order to assess the autonomy of weapon systems, there is a prevalent framework in the literature on autonomous weapon systems. This framework exemplifies three types of weapon systems in relation to the role played by human operators in the targeting loop. According to this framework, human operators can either be *in-the-loop*, *on-the-loop*, or *out-of-the-loop*. Human-in-the-loop systems are those systems that have autonomy for some tasks (e.g., tracking, detecting, prioritizing targets), but they depend on the human operator to make the ultimate decision of firing to a particular target. Human on-the-loop systems are considered to execute all the tasks in the targeting loop independently of human operators, but they cannot finalize the targeting loop without the approval of a human operator. This means that human operators can also intervene and override the system's actions. According to DoD (2023), human-on-the-loop systems are also considered autonomous weapon systems because they can undergo the targeting cycle on their own, that is, independently of human operators. The third type of system is the human out-of-the-loop system. This type of system is the focus of the debate on LAWS because these systems require neither human input, as in the human in the loop, nor supervision, as in the human on the loop systems. The moral question raised in this thesis also mainly concerns this latter type of weapon system.

Consequently, the analysis in Chapter 2 aims to clarify the confusion around the definitions of LAWS. I conclude this chapter by pointing out that the lethality of these systems is an important factor, which explains my deliberate use of the term "lethal autonomous weapon systems" instead of, for example, autonomous weapon systems.

Chapter 3 concerns the ethical issues surrounding LAWS. The chapter starts with a brief explanation of the ethical theory that governs the conduct in warfare: just war theory, more precisely, *jus in bello,* and legal framework, international humanitarian law. However, the main ethical discussion in this chapter revolves around particular principles of *jus in bello* and IHL: the principle of distinction and the principle of proportionality. The former principle prohibits attacks on noncombatants, and the latter prohibits attacks on civilians that are not proportionate to the military advantage to be gained.

Critics of LAWS claim that LAWS will fall short of complying with these two principles because compliance with these principles requires human judgment. For instance, the category of civilian and combatant is unclear because it often depends on the awareness of the humans' behavior at a certain time. The obscurity in defining the categories of who is a legitimate target is often resolved by human situational awareness. For example, soldiers who fall sick, wounded, or surrendering are considered hors de combat; they are no longer legitimate targets. Similarly, the principle of proportionality is a challenge that LAWS would have difficulty abiding by because it requires the judgment of how much collateral damage is acceptable for the military advantage to be gained. The proportionality principle poses a problem because LAWS would have to make decisions on the lives of innocent victims. This implies that the decisions, such as the number of civilians casualties resulting from an attack, would be made by machines.

Another factor, critics argue, that makes LAWS inappropriate for warfare is the inherent unpredictability in LAWS. Unpredictability is an important concern because it would constitute the reason for a ban on LAWS. After all, unpredictability in the war would mean that LAWS cannot be entrusted for compliance with the principles of *just in bello*. That is, the inherent unpredictable behavior of LAWS would make their deployment unethical because humans cannot predict if their action would comply with the principles of distinction and proportionality. Therefore, there would inevitably be violations of *jus in bello* principles.

Contrary to the opponents' views, proponents of LAWS contend that these machines can (potentially) reduce the unpredictability of warfare by reducing human errors. They highlight that LAWS' advanced sensory abilities outperform the human soldiers' capabilities. LAWS' capacity to process data faster than humans could lead to minimizing human errors in warfare. This means that LAWS can collect more information before taking lethal action. Additionally, deploying LAWS would prevent human soldiers from harm's way, mitigating the risk of casualties. As a result, LAWS would reduce unpredictability and unethical behavior in warfare, as they would not be subject to the physical, psychological, and cognitive limitations that humans have.

Proponents argue that these systems will have relatively less unpredictable behavior than humans because they will be based on rule-based algorithms. According to this, LAWS' autonomy means that they operate without direct human input, but they strictly follow instructions given beforehand to the system's algorithm. For this reason, LAWS would perform better than human soldiers because they will be bound by strict ethical rules that they will abide by in most circumstances. Since the unpredictability of LAWS shapes the views in favor of and against the use of LAWS, it has been discussed in relation to its ethical implications.

In Chapter 4, I have focused on the question of whether the use of LAWS leads to a responsibility gap. The responsibility gap stems from the concern that the moral harm caused by LAWS is not attributable to anyone. In this chapter, I first analyze moral responsibility and then proceed to the argument of responsibility gap. The responsibility gap argument relies on the premise that no one has direct control over the actions of LAWS, and the machine itself cannot be held responsible. Therefore, there is a moral responsibility gap.

As a solution to the responsibility gap problem, I proposed that the responsibility is not always assigned on the condition of the presence of direct control. To show that direct control is not always required for moral responsibility assignments, I have presented vicarious responsibility, where an agent can be held responsible for an action of another. Vicarious responsibility is an obscure concept because it aims to connect an agent to another's actions, even if there is no direct link or relation between these two agents. To overcome the obscurity inherent in vicarious responsibility, I have used a modified version of a formal definition of vicarious responsibility proposed by Glavanicova & Pascucci (2022) and applied the definition to the case of LAWS. Following this definition, I concluded that LAWS designers can be held morally responsible for the moral harm caused by LAWS because of their unique moral relation with their creation. At the end of the chapter, I have also discussed that the moral responsibility of LAWS' designers in a vicarious sense is aligned with other solutions to the moral responsibility gaps: collective and distributed responsibility. Therefore, the vicarious responsibility of designers of LAWS overcomes the problem of the responsibility gap in the context of LAWS.

BIBLIOGRAPHY

Aristote, Ross, W. D., & Brown, L. (2009). *The Nicomachean ethics* (ed. rev). Oxford University Press.

Alpaydin, E. (2016). *Machine learning: The new AI*. MIT Press.

Altmann, J., Sauer, F. (2017). Autonomous weapon systems and strategic stability. *Survival, 59*(5), pp. 117-142

Anderson, K., & Waxman, M. C. (2013). Law and ethics for autonomous weapon systems: Why a ban won't work and how the laws of war can. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2250126

Anderson, K., & Waxman, M. C. (2017). *Debating Autonomous Weapon Systems, their Ethics, and their Regulation under International Law* (R. Brownsword, E. Scotford, & K. Yeung, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199680832.013.33

Arkin, R. C. (2009). *Governing Lethal Behavior in Autonomous Robots*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420085952

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics, 9*(4), pp. 332-341. https://doi.org/10.1080/15027570.2010.536402

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

Bartneck et al. (2021). Military uses of AI. In *An Introduction to Ethics in Robotics and AI.* SpringerBriefs in Ethics. https://doi.org/10.1007/978-3-030-51110-4_11

Beard, J. M. (2018). The principle of proportionality in an era of high technology. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3119384

Beernaert, T. F., Bayrak, A. E., Etman L. F. P., & Papalambros, P. Y. (2018). Framing the concept of autonomy in system design. Conference Paper. 15th International Design Conference, pp. 2821-2832. https://doi.org/10.21278/idc.2018.0281

Beavers, G. & Hexmoor, H. (2004). Types and limit of agent autonomy. In M. Nickles, M. Rovatsos, & G. Weiss (Eds.), Agent and computational autonomy: Potential, risks, and solutions, pp. 95-103. Springer.

Birnbacher, D. (2016). Are autonomous weapons systems a threat to human dignity? In N. Bhuta, S. Beck, R. Geiβ, H. Liu, & C. Kreβ (Eds.), *Autonomous Weapons Systems: Law, Ethics, Policy* (pp. 105-121). Cambridge: Cambridge University Press. doi:10.1017/CBO9781316597873.005

Bradshaw, J. M., Hoffman, R. R., Johnson, M., & Woods, D. D. (2013). The seven deadly myths of "Autonomous Systems." *IEEE Intelligent Systems*, 28(3), pp. 54–61. https://doi.org/10.1109/MIS.2013.70

Bringsjord, S. & Govindarajulu, N. S. (2022). Artificial Intelligence. In E.N. Zalta & U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (fall 2022 Edition). Retrieved from https://plato.stanford.edu/entries/artificial-intelligence/

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1),pp. 1-12. https://doi.org/10.1177/2053951715622512

Boulanin, V., & Verbruggem, M. (2017). *Mapping the development of autonomy in weapon systems*. Stockholm International Peace Research Institute. https://www.sipri.org/publications/2017/other-publications/mapping-development-autonomy-weapon-systems

Crootof, R. (2015). The killer robots are here: Legal and policy implications. *Cardozo Law Review, 36,* pp. 1837-1915. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2534567

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299–309. https://doi.org/10.1007/s10676-016-9403-3

Defense Advanced Research Projects Agency (n.d.). *About DARPA. https://www.darpa.mil/about-us/about-darpa.*

De Jong, R. (2020). The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm. *Science and Engineering Ethics*, *26*(2), 727–735. https://doi.org/10.1007/s11948-019-00120-4

De Vries, B. (2023). *Individual Criminal Responsibility for Autonomous Weapons Systems in International Criminal Law.* Koninklijke Brill.

Diakopoulos, N. (2020). Transparency. In Dubber, M. D., Pasquale, F., & Das, S. (Eds.). *The Oxford handbook of ethics of AI*. Oxford University Press.

Dumouchel, P. (2021). Lethal autonomous weapons systems: Organizational and political consequences. *Philosophical Journal of Conflict and Violence*, *5*(1). https://doi.org/10.22618/tp.pjcv.20215.1.139006

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books, a member of the Perseus Books Group.

Ezenkwu, C. P., & Starkey, A. (2019). Machine Autonomy: Definition, Approaches, Challenges and Research Gaps. In K. Arai, R. Bhatia, & S. Kapoor (Eds.), *Intelligent Computing* (Vol. 997, pp. 335–358). Springer International Publishing. https://doi.org/10.1007/978-3-030-22871-2_24

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society, 374*(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Future of Life Institute (2016). *Autonomous Weapons: An Open Letter from AI & Robotics Researchers,* http://futureoflife.org/open-letter-autonomous-weapons/

Future of Life Institute (2018). *Lethal Autonomous Weapon Pledge: An Open Letter*. https://futureoflife.org/open-letter/lethal-autonomous-weapons-pledge/

Future of Life Institute (n.d.). *Slaughterbots are already here*. https://futureoflife.org/project/lethal-autonomous-weapons-systems/

Galliott, J., & Forge, J. (2021). Debate on the ethics of developing AI for lethal autonomous weapons. *Philosophical Journal of Conflict and Violence*, *5*(1). https://doi.org/10.22618/tp.pjcv.20215.1.139009

Gotterbarn, D. (2001). Informatics and professional responsibility. *Science and Engineering Ethics*, *7*(2), 221–230. https://doi.org/10.1007/s11948-001-0043-5

Glavanicova, D. & Pascucci, M. (2022). Making sense of vicarious responsibility: Moral philosophy meets legal theory. *Erkenntnis. https://doi.org/10.1007/s10670-022-00525-x*

Glavanicova, D. & Pascucci, M. (2022). Vicarious liability: A solution to a problem of AI responsibility? *Ethics and Information Technology, 24*. p. 28. https://doi.org/10.1007/s10676-022-09657-8

Goetze, T. S. (2021). Moral Entanglement: Taking Responsibility and Vicarious Responsibility. *The Monist*, *104*(2), 210–223. https://doi.org/10.1093/monist/onaa033

Goetze, T. S. (2022). Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 390–400. https://doi.org/10.1145/3531146.3533106

Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology*, *15*(2), 99–107. https://doi.org/10.1007/s10676-012-9301-2

Henckaerts, J.-M., Doswald-Beck, L. (2009). *Customary international humanitarian law*. C. Alvermann & International Committee of the Red Cross (Eds.). Cambridge University Press.

Heyns, C. (2013). Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, United Nations Doc. A/HRC/23/47

Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice, 22*(30), pp. 731-747. https://www.jstor.org/stable/45217332

Holland Michel, A. (2020). The Black Box, Unlocked: Predictability and Understandability in Military AI. Geneva, Switzerland: United Nations Institute for Disarmament Research. doi:10.37559/SecTec/20/AI1

Huang, H.M., Messina, E., Wade, R., English, R., Novak, B., & Albus, J. (2004). Autonomy measures for robots. Dynamic Systems and Control. Conference paper. 1241–1247. https://doi.org/10.1115/IMECE2004-61812

Human Rights Watch (2012). Losing humanity: The case against killer robots. https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots

Human Rights Watch (2016), Making the case: the dangers of killer robots and the need for a preemptive ban. https://www.hrw.org/report/2016/12/09/making-case/dangers-killer-robots-and-need-preemptive-ban

International Committee of Red Cross (2016). Views of the ICRC on autonomous weapon systems, https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system

International Committee of Red Cross (n.d.) *International humanitarian law treaties: essential documents*. Online. https://blogs.icrc.org/cross-files/international-humanitarian-law-treaties-essential-documents/ (last accessed 30 August 2023).

International Committee of Red Cross (n.d.). *Customary law*. https://www.icrc.org/en/war-and-law/treaties-customary-law/customary-law#:~:text=Treaties%2C%20such%20as%20the%20four,States%20formally%20establish%20certain%20rules. (last accessed 30 August 2023)

Israel Aerospace Industries (n.d.). *Harpy*. https://www.iai.co.il/p/harpy (Last accessed 30 August 2023).

Johnson, D. G. & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology, 7*(2), 99-107. https://doi.org/10.1007/s10676-005-4585-0

Johansson, L. (2018). Ethical aspects of military maritime and aerial autonomous systems. *Journal of Military Ethics*, *17*(2–3), 140-155. https://doi.org/10.1080/15027570.2018.1552512

Jotterand, F., & Bosco, C. (2020). Keeping the "Human in the loop" in the age of artificial intelligence. *Science and Engineering Ethics*, *26*(5),2455-2460. https://doi.org/10.1007/s11948-020-00241-1

Klincewicz, M. (2015). Autonomous weapons systems, the frame problem and computer security. *Journal of Military Ethics, 14*(2). pp. doi:162-176. 10.1080/15027570.2015.1069013

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology*, *24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Lele, A. (2017). A military perspective on lethal autonomous weapon systems. In *Perspectives on Lethal Autonomous Weapon Systems*. UNODA Occasional Papers No. 30, New York: United Nations Publications, pp. 57-68

Leveringhaus, A. (2016). *Ethics and Autonomous Weapons* (1st ed). Palgrave Pivot. London

Levy, N. (2014). *Consciousness and moral responsibility* (1st ed). Oxford University Press.

Matthias, A. (2004). The responsibility gap in ascribing responsibility for the actions of automata,' *Ethics and Information Technology 6*(3), 175–183

McFarland, T. (2015). Factors shaping the legal implications of increasingly autonomous military systems. *International Review of the Red Cross, 97*(900), 1313–1339

McFarland, T. (2020). *Autonomous Weapon Systems and the Law of Armed Conflict: Compatibility with International Humanitarian Law*. Cambridge Press

Mele, A., & Sverdlik, S. (1996). Intention, intentional action, and moral responsibility. *Philosophical Studies*, *82*(3), 265–287. https://doi.org/10.1007/BF00355310

Mellor, R. (2021). Faces of Vicarious Responsibility. The Monist, 104(2), 238–250. https://doi.org/10.1093/monist/onaa035

Melzer, N., Kuster, E. (2016). International humanitarian law: a comprehensive introduction. Switzerland: ICRC.

Miller, S. (2019). Machine Learning, ethics and law. *Australasian Journal of Information Systems*, *23*. https://doi.org/10.3127/ajis.v23i0.1893

Nissenbaum, H. (1996). Accountability in a computerized society. *Sciences and Engineering Ethics, 2*(1), 25-42

Noorman, M., Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology, 16*, 51-62

Noorman, M. (2023). Computing and moral responsibility. In E. N. Zalta & U. Nodelman (eds.). *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition). Retrieved from https://plato.stanford.edu/entries/computing-responsibility/

Noone, G. R., Noone, D. C. (2015). Debate over autonomous weapons systems. *Case Western Reserve Journal of International Law 47*(1), 25-35

Organisation for Prohibition of Chemical Weapons. (n.d.). *History*. https://www.opcw.org/about-us/history

Osinga, F. (2005). *Science, strategy and war: The strategic theory of John Boyd*. Eburon Academic Publishers.

Oxford Learner's Dictionaries. (n.d.). Autonomy. In oxfordlearnersdictionaries.com. Retrieved from https://www.oxfordlearnersdictionaries.com/definition/english/autonomy?q=autonomy

Pagallo, U. (2017). When morals ain't enough: Robots, ethics, and the rules of the law. *Minds and Machines*, *27*(4), 625-638. https://doi.org/10.1007/s11023-017-9418-5

Picard, R. W. (2000). *Affective computing*. The MIT Press

Santoni De Sio, F. & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*(15). doi:10.3389/frobt.2018.00015

Santoni De Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Scharre, P. (2020). *Autonomous Weapons and Stability* [Doctoral thesis, King's College London]. King's Research Portal. https://kclpure.kcl.ac.uk/ws/portalfiles/portal/129451536/2020_Scharre_Paul_1575997_ethesis.pdf

Scharre, P., & Horowitz, M. C. (2015). *An Introduction to autonomy in weapon systems*. Center for a New American Security. http://www.jstor.org/stable/resrep06106

Schmitt, M. N., Thurnher, J. S. (2013). Out of the loop: Autonomous weapon systems and the law of armed conflict. *Harvard National Security Journal, 4*, 231-281

Schulzke, M. (2013). Autonomous Weapons and Distributed Responsibility. *Philosophy & Technology*, *26*(2), 203–219. https://doi.org/10.1007/s13347-012-0089-0

Schwitzgebel, E. & Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest Studies in Philosophy, 39*(1). pp. 98-119.

Sharkey, A. (2019). Autonomous weapons systems, killer robots and human dignity. *Ethics and Information Technology*, *21*(2), 75-87. https://doi.org/10.1007/s10676-018-9494-0

Sharkey, N. (2012). The evitability of autonomous robot warfare,*' International Review of the Red Cross, 94*(886), pp.787- 799

Sharkey, N. (2017). Why robots should not be delegated with the decision to kill. *Connection Science, 29*(2), 177-186.

Smithers, T. (1997). Autonomy in Robots and Other Agents. *Brain and Cognition*, *34*(1), 88–106. https://doi.org/10.1006/brcg.1997.0908

Simpson, T. W., & Müller, V. C. (2016). Just war and robots' killings. *Philosophical Quarterly*, *66*(263), 302-322. https://doi.org/10.1093/pq/pqv075

Snowdon, P. & Gomes, A. (2023). Peter Frederic Strawson. In E. n. Zalta & U. Nodelman (eds.). *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition). Retrieved from https://plato.stanford.edu/entries/strawson/

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62-77

Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics and International Affairs*, *30*(1), 93-116. https://doi.org/10.1017/S0892679415000647

Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.

Stop Killer Robots (n.d.). *New international law is needed.* https://www.stopkillerrobots.org/stop-killer-robots/we-can-stop-killer-robots/ (Last accessed 30 August)

Taddeo, M. & Blanchard, A. (2022a). Accepting moral responsibility for the actions of autonomous weapon systems- a moral gambit. *Philosophy & Technology, 35*(78). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2534567

Taddeo, M. & Blanchard, A. (2022b). A comparative analysis of the definitions of autonomous weapon systems. *Science and Engineering Ethics, 28*(37). https://doi.org/10.1007/s11948-022-00392-3

Talbert, M. (2022). Moral responsibility. In E. N. Zalta & U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition). Retrieved from https://plato.stanford.edu/entries/moral-responsibility/

Taylor, I. (2020). Who Is Responsible for Killer Robots? Autonomous Weapons, Group Agency, and the Military-Industrial Complex. *Journal of Applied Philosophy*, *38*(2), 320–334. https://doi.org/10.1111/japp.12469

Trusilo, D. (2023). Autonomous AI systems in conflict: emergent behavior and its impact on predictability and reliability. Journal of Military Ethics. https://doi.org/10.1080/15027570.2023.2213985

US Department of Defense. (2023). *Department of Defense Directive: Autonomy in Weapon Systems*. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf

Van De Poel, I., Royakkers, L., & Zwart, S. D. (2015). *Moral Responsibility and the Problem of Many Hands* (1 ed.). Routledge. https://doi.org/10.4324/9781315734217

Vilmer, J. J. (2015). *Terminator Ethics: Should We Ban "Killer Robots"?* Ethics and International Affairs. https://www.ethicsandinternationalaffairs.org/online-exclusives/terminator-ethics-should-we-ban-killer-robots

Wallach, W. & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

APPENDICES

A. TURKISH SUMMARY / TÜRKÇE ÖZET

Bölüm 1: Giriş

Ölümcül Otonom Silah Sistemleri (OSS), yapay zekanın askeri kullanımına ilişkin etik tartışmalarda önemli bir ilgi görmüştür. Bununla birlikte, OSS'leri modern savaşta zaten yaygın olarak kullanılan uzaktan kumandalı insansız hava araçlarından ayırmak gerekir. İnsansız hava araçları, benzer teknolojik özelliklere sahip olsalar da ölümcül otonom silah sistemleri olarak kabul edilmezler. İnsansız hava araçlarının otonom olarak hedefleri seçip saldırıda bulunmazlar. İnsansız hava araçları saldırı eylemleri gerçekleştirmek için insan operatörlere ihtiyaç duymaktadır. Uzaktan kumandalı sistemler 'döngünün içerisinde insan' sistemleri olarak adlandırılabilinir çünkü insan operatörler özellikle ölümcül kararlarda önemli bir rol oynar. Buna karşın, OSS ölümcül saldırının başlatılmasında insan yargısına olan ihtiyacı ortadan kaldırır. Dolayısıyla, OSS, eylemlerinin arkasında doğrudan insan kontrolü olmadan çalışabilme kabiliyetleri bakımından insansız hava araçlarından farklıdır.

Aktive edildikten sonra, OSS'lerin eylemleri artık bir insan operatörün doğrudan kontrolüne veya denetimine bağlı olmayacaktır. Uzaktan kumanda yoluyla bazı işlevleri yerine getirebilen mevcut sistemlerden farklı olarak yapay zeka, OSS'lerin çevresine uyum sağlama, çevresinden öğrenme ve insan müdahalesi olmadan hedefleri tespit etme kabiliyetine sahip olmasını sağlar.

Öte yandan, bu yetenekler dezavantajları da beraberinde getirmektedir. Yapay zeka, özellikle de makine öğrenmesi, geleneksel programlama tekniklerinden farklıdır. Geleneksel programlamada girdi, programcı tarafından tanımlanan sabit bir algoritma kullanılarak işlenir. Böylece, hangi girdinin hangi çıktı ile sonuçlanacağı programcı tarafından bilinir. Buna karşılık, makine öğrenmesi ile donatılmış sistemlere çok miktarda veri sağlanır ve sistem algoritmasını eğitim verilerinden üretir.

Sistem algoritmayı ürettiği için, programcı girdiyi çıktıya dönüştürme prosedürünü tam olarak bilemeyebilir Bu şeffaflık eksikliği, makine öğrenimi sistemlerinin doğasında var olan karmaşıklık nedeniyle bazı durumlarda sistemin çıktısını tahmin etmenin zorlaştığı anlamına gelir.

Makine öğrenimi ile donatılmış yeni sistemlerin bu özelliği, yani öngörülemeyen davranışları, OSS'lerin savaşta kullanılmasına karşı temel bir sorun olarak algılanmaktadır. Bu nedenle aktivistler, STK'lar ve akademisyenler, OSS'lerin uluslararası düzeyde yasaklanması çağrısında bulunmaktadır. Bu sistemlerin kullanılması hem etik dışı hem de hukuka aykırı olduğunu savunmaktadırlar. OSS'lerin kullanılması ilgili üç önemli sorun dikkat çekmektedir: ayrım ilkesi, orantılılık ilkesi ve sorumluluktaki boşluk. Bu nedenle, bu tezde, OSS'lerin kullanımının oluşturacağı etik sorunları analiz edeceğim. Bu sistemlerin Uluslararası İnsancıl Hukuk ve haklı savaş teorisi ilkelerine uygunluğuna ilişkin ön değerlendirmelerin ardından, asıl tezin sorusu olan sorumluluk boşluğu sorununa yer verilmiştir. Ahlaki sorumluluk önemli bir faktördür çünkü makine ölüm kalım kararları verecekse, bu kararların sonuçlarından kimin sorumlu olacağı önem kazanır.

Bölüm 2: Ölümcül Otonom Silah Sistemleri(OSS) Nedir?

Bu bölümde, OSS'lerde otonomi kavramını anlamak için kullanılan çeşitli tanımları ve çerçeveleri analiz edeceğim. Bu analiz, ABD Savunma Bakanlığı (DoD) ve Uluslararası Kızıl Haç Komitesi (ICRC) tarafından sağlanan ve sıklıkla atıfta bulunulan OSS tanımlarını içermektedir. Bu tanımlara ek olarak, bu bölüm otonom silah sistemlerini insan müdahalesi çeşitlerine göre sınıflandıran ve yaygın olarak kullanılan döngü çerçevesini de incelemektedir: döngü içinde insan, döngü üzerinde insan ve döngü dışında insan. Tez, özellikle karar verme prosedürünün tamamen sistemin algoritmasına devredildiği döngü-dışında-insan kategorisindeki sistemlere odaklanmaktadır.

ABD Savunma Bakanlığı (DoD) 2012 tarihli Direktifinde (2023'te güncellenmiştir) otonom silah sistemini "bir kez etkinleştirildiğinde, bir operatörün daha fazla müdahalesi olmaksızın hedefleri seçebilen ve bunlara saldırabilen bir silah sistemi" olarak tanımlamaktadır (2023, s. 21). DoD'ye göre, bir insan operatörün müdahalesinden bağımsızlık, bu sistemlerin tanımlanmasında çok önemli bir rol oynamaktadır. DoD'nin tanımı literatürde en çok öne çıkan tanımlardan biri olmakla birlikte, bu sistemlerin ne olduğu tartışılırken kafa karıştırıcı da olabilir. İnsan operatörlerden bağımsızlık, ölümcül otonom silah sistemlerinin çok önemli bir parçasıdır. Ancak bu sistemleri sadece operatörlerden bağımsız olarak hareket edebilen sistemler olarak tanımlamak yeterli değildir. Bu tanım ilk bakışta açıklayıcı görünmekle birlikte, hedef seçme ve hedefe angaje olma eylemlerinin neleri kapsadığını açıklamakta yetersiz kalmaktadır. Başka bir deyişle, OSS'lerin insan operatörler olmadan ne yaptığı açık değildir.

Yaygın olarak atıfta bulunulan bir diğer OSS tanımı Uluslararası Kızıl Haç Komitesi (ICRC) tarafından yapılmıştır. ICRC, OSS'leri "kritik işlevlerinde otonomiye sahip herhangi bir silah sistemi. Yani, insan müdahalesi olmadan hedefleri seçebilen (arama veya tespit etme, tanımlama, izleme, seçme) ve saldırabilen (yani, hedeflere karşı güç kullanma, etkisiz hale getirme, hasar verme veya yok etme) bir silah sistemi" olarak tanımlamaktadır (ICRC, 2016, s. 1). Bu iki tanım birbirini tamamlar niteliktedir, zira her ikisi de OSS'lerin insan operatörünün müdahalesinden bağımsız olma özelliğini vurgulamaktadır. Ayrıca ICRC, bu sistemler tarafından yerine getirilen görevleri detaylandırarak DoD'nin tanımındaki boşluğu da doldurmaktadır.

2.1. Döngü Çerçevesi

OSS'leri tanımlarken kullanılan en yaygın yöntem, döngü üzerinden bu sistemleri kategorize etmektir. Bu döngü çerçevesine göre üç tür otonom silah sisteminden bahsetmek mümkündür. Bu sistemler şu şekildedir: (i) İnsan-döngünün-içinde; (ii) insan-döngünün-üzerinde; (iii) insan-döngünün-dışında.

Birincisi, silah sisteminin hedefleme döngüsünü yalnızca operatörün aktif olarak angaje edilecek hedefleri seçmesi durumunda gerçekleştirebileceğini belirten döngüde insan. Döngü içinde insan silah sistemleri yarı-otonom silah sistemleri olarak da bilinir. Operatör hedef seçimi kararı üzerindeki kontrolünü sürdürür. Bu kategoriye giren silah sistemlerine örnek olarak ateşle-ve-unut mühimmatları verilebilinir. Ateşle-ve-unut sistemler operatörün seçtiği hedefi vurma ihtimali artıracak, ya da hareket halinde olan hedefi vurmaya yardımcı olacak araçlarla donatılmışlardır.

İkincisi, insan denetimli veya operatör denetimli otonom silah sistemleri olarak da adlandırılan döngü üzerinde insan sistemleridir. Bu tür sistemlerde operatörün rolü sistemin faaliyetini izlemektir. Ancak operatör müdahale etmezse, sistem döngüdeki görevleri bağımsız olarak kendi başına yerine getirecektir. Döngüde insan bulunan silah sistemlerine bir örnek, İsrail Havacılık ve Uzay Sanayii (IAI) tarafından geliştirilen Harpy'dir. Harpy, fırlatıldıktan sonra, düşman radarlarını aramak için önceden tanımlanmış bir alanda dolaşır. Tespit edildiğinde radarları vurur ve imha eder.

Son olarak ise döngünün-dışında-insan silah sistemleridir. Bu tür bir sistem OSS'lerin genel imajını en çok yansıtan sistemdir. Bu tür sistemler döngü içinde insan sistemlerinde olduğu gibi herhangi bir kontrol ya da döngü üzerinde insan sistemlerinde olduğu gibi bir operatörün denetimi olmaksızın bağımsız olarak hareket eder. Bu tür sistemler operatörün müdahale etme veya operasyonu durdurma yeteneği olmadan hedefleri vurabilen sistemlerdir. Önemli bir örnek Güney Koreli DODAAM şirketi tarafından geliştirilen Super aEgis II adlı robotik nöbetçi sistemidir. Otonom modda Super aEgis II, potansiyel bir hedefin ısısını ve hareketini tespit etmek için termal sensörler ve kameralar kullanır ve bu girdi verilerine dayanarak insan kontrolü veya denetimi olmadan hedefleme kararları verebilir.

Bölüm 3: OSS'leri çevreleyen etik sorunlar

Haklı savaş teorisi, ne zaman savaşa girileceğine (jus ad bellum) ve savaşta etik olarak kabul edilebilir davranışlara (jus in bello) ilişkin kriterler sağlayan etik çerçevedir. Altı kriterin karşılanması halinde savaşa girmek haklı görülebilir: "haklı sebep, orantılılık, gereklilik, son çare, doğru otorite ve makul başarı olasılığı" (Leveringhaus, 2016, s. 12). Öte yandan, savaşta etik davranış üç kriterden oluşmaktadır: "ayrım, araçların orantılılığı ve gereklilik" (Leveringhaus, 2016, s. 12). Jus in bello ilkeleri OSS'ler üzerinde olan tartışma ile daha ilgilidir, çünkü jus ad bellum savaşa girmenin haklı olup olmadığı gibi sorularla ilgilidir. Jus in bello ise kimin meşru hedef olduğu, savaşta hangi araçların meşru araçlar olduğu vb. sorularla ilgilidir. Dolayısıyla, etik tartışma daha çok OSS'lerin jus in bello ilkelerine uyup uyamayacağına odaklanmaktadır.

Jus in bello ve Uluslararası İnsancıl Hukuk(UIH)'un iki temel ilkesi, OSS'ler ile en ilgili ilkeler olarak kabul edilmektedir: ayrım ilkesi ve orantılılık ilkesi. Ayırt etme ilkesi, çatışmanın taraflarını "sivil halk ile savaşçılar arasında ve sivil nesneler ile askeri hedefler arasında ayrım yapmaya ve ... operasyonlarını yalnızca askeri hedeflere yöneltmeye" yasal olarak zorlar (ICRC, AP I, Md. 48).

Ölümcül otonom silah sistemlerini eleştirenler, bu sistemlerin ayrım ilkesine uyma konusunda yetersiz olma ihtimalinin yüksek olduğunu iddia etmektedir. Sparrow'a göre bu sistemler savaşanları sivillerden ayırt edemez çünkü bu sistemlerin eylemleri öngörülemezdir (2007). Benzer şekilde Asaro, insan zekâsıyla karşılaştırıldığında, OSS'lerin "en iyi ihtimalle öğrenme ve adaptasyon için yalnızca son derece sınırlı yeteneklere sahip olacağını, savaşın sisi ile başa çıkabilecek sistemler tasarlamanın zor ya da imkânsız olacağını” savunmaktadır (2012, s. 692). Asaro (2012) ayrıca savaş alanının karmaşıklığının, özellikle bu makinelerin ayrım ilkesine uyma kabiliyeti açısından, askeri robotikçilerin beklentilerini aştığını iddia etmektedir. Bilgisayar bilimcisi Noel Sharkey, ayrım ilkesi için gerekli olan üç hususu vurgulayarak bu düşünceye katılmaktadır. İlk olarak, bu sistemlerin askerleri sivillerden ayırmak için yeterli “duyusal veya görsel işleme sistemlerinden” yoksun olduğunu iddia etmektedir (Sharkey, 2017, s. 179). İkinci olarak, sivilin net bir tanımının olmaması, sivilleri ayırt edecek bir program kodlama girişimi için başlı başına bir zorluktur çünkü siviller UIH’a göre “muharip olmayan kişi” olarak tanımlanmaktadır (Sharkey, 2017, s. 179; Sharkey, 2012, s. 789). Son olarak, Sharkey, ayrım ilkesine uyulmasını sağlamak için insana özgü yargı yeteneğinin gerekli olduğunu, çünkü görme işleminin ayrım kararları vermek için yetersiz olduğunu savunmaktadır. Yani, duyusal teknolojiler ileri bir seviyeye ulaşsa bile, makineler yine de “ayrımcılık kararlarına yardımcı olacak savaş alanı farkındalığına veya sağduyulu muhakemeye” sahip olmayacaktır (Sharkey, 2017, s. 179).

Bu iddialarını değerlendirmek için, bir bireyin muharip veya muharip olmayan olarak sınıflandırılmasını belirleyen koşulları ve savaşta bu statülerin nasıl atandığını incelemek faydalı olacaktır. Bir kişinin muharip mi yoksa sivil mi olduğunu belirlerken, üniforma giymek genellikle ilk koşullardan biri olarak kabul edilir. Üniformalar bireylerin silahlı bir gücün üyeleri olarak tanımlanmasına yardımcı olur. Ancak bazı durumlarda üniforma giyen kişiler bile “hors de combat” olarak kabul edilir, yani artık meşru hedef değildirler. Yaraları veya hastalıkları nedeniyle çatışmalara katılmaya devam edemeyecek durumda olan veya teslim olma niyetini açıkça ifade eden savaşçılar meşru askeri hedefler olarak hedef alınamazlar. Teslim olma niyetini tespit etmek daha da önemli bir zorluk teşkil eder, çünkü bir askerin teslim olma niyetini açıkça ifade edebileceği yollar, beyaz bayrak göstermek, sözlü iletişim, her iki kolunu kaldırmak veya silahları yere bırakmak gibi önemli ölçüde farklılık gösterebilir. Sonuç olarak, bu tür durumlarda, bir kişinin belirli bir üniforma giyip giymediğini tespit eden duyu ve görüş işleme sistemleri bile o hedeflere saldırmak için yeterli olmayacaktır. Sistem ayrıca üniforma giyen kişinin hastalık, yaralanma ya da teslim olma niyeti nedeniyle savaş dışı statüsünde olup olmadığını da tespit etmelidir.

Üniforma giyen muhariplerin belirlenmesiyle ilgili bir başka sorun da modern savaşın doğasıdır. Günümüzde silahlı çatışmaların değişen doğası göz önüne alındığında, üniforma giyme kriteri sadece bazı durumlarda yeterli olabilir. Örneğin, bir devletin bir milisle silahlı çatışmaya girdiği uluslararası olmayan çatışmalarda, isyancı gruplar ayırt edici bir üniforma giymeyebileceğinden, muharipler ile siviller arasında ayrım yapmak özellikle zorlaşır. Tüfek gibi bir silah taşımanın o kişiyi tanımlamak ve hedef almak için yeterli olacağı iddia edilebilir. Örneğin, duyu sistemleri askeri hedefleri seçmek için tüfekleri veya askeri teçhizatı tespit edebilir. Ancak uluslararası silahlı çatışmalarda karşılaşılan zorluklar bu durumlarda da geçerlidir; yani sistem tüfek taşıyan kişinin yaralı, bilinçsiz ya da teslim olmuş olup olmadığını, dolayısıyla savaşa aktif olarak katılmadığını da tespit etmelidir. Ayrıca, makine ölülerini gömen isyancıları veya tüfek taşımaya zorlanan çocukları da tanımalıdır. Bu hususlar göz önüne alındığında, ayrım ilkesine uyum iki yönlü bir zorluk teşkil etmektedir. İlk olarak, bir kişinin düşman askeri mi yoksa savaşçı olmayan biri mi olduğunun tespit edilmesini gerektirir. Dahası, eğer asker olduğu tespit edilirse, hedefin savaş dışı (hors de combat) statüsüne sahip olup olmadığı da tespit edilmelidir. OSS'leri eleştirenlere göre bu zorlukların algoritmalarla aşılması mümkün değildir.

Benzer şekilde, eleştirmenler OSS'lerin jus in bello'nun orantılılık ilkesine uymakta yetersiz kalacağını iddia etmektedir. Orantılılık ilkesi, "öngörülen somut ve doğrudan askeri avantaja kıyasla aşırı olacak şekilde, sivil can kaybına, sivillerin yaralanmasına, sivil nesnelerin zarar görmesine veya bunların bir kombinasyonuna neden olması beklenebilecek" saldırıları yasaklar (ICRC, AP I, Md. 51). Dolayısıyla, sivillerin veya sivil nesnelerin kasıtlı olarak hedef alınması meşru değildir. Ancak, askeri hedeflere saldırırken belli ölçülerde sivil kayıplardan kaçınılamıyorsa, askeri hedeflere yönelik kasıtlı saldırıların ikincil hasarı veya yan etkileri olarak bu saldırılar meşru kabul edilirler. Orantılılık ilkesi, bir çatışmanın taraflarına, elde edilecek askeri avantajın bir yan etkisi olarak aşırı ikincil hasara neden olmamalarını emreder. Sharkey (2017) iki tür orantılılık olduğunu savunmaktadır: kolay ve zor orantılılık. Ona göre, OSS'ler yalnızca kolay orantılılık ile baş edebilirler; yani makineler en uygun silahı veya mühimmatı seçerek ve uygun şekilde yönlendirerek ikincil zararı azaltmaya yardımcı olabilir. Örneğin, hassas güdümlü mühimmatlar, yerleşik sensörleri ile daha doğru hedefleme sağlayarak ayrım gözetmeyen ve orantısız saldırıları azaltmıştır. Benzer bir şekilde, OSS'ler, farklı saldırı seçenekleri arasından askeri avantajı korurken ikincil hasarı görece azaltan en uygun mühimmatı seçerek orantısız saldırıları azaltmaya yardımcı olabilir. Ancak Sharkey, makinelerin zor orantılılık kararları veremeyeceğini, yani elde edilecek askeri avantaj için en başta ölümcül güç uygulanıp uygulanmayacağına makinaların karar veremeyeceğini iddia eder (2017).

Dolayısıyla, savaş alanında OSS kullanarak amaçlanan bir saldırının ikincil zararını en aza indirerek kolay orantılılık mümkün olabilir. Ancak, karmaşık bir orantılılık durumunda ölümcül güç uygulama kararını makinelerin hesaplaması zor olmaya devam edecektir. Hesaplayamazlar çünkü orantılılık, potansiyel çağrışımlarının aksine sadece sayısal bir hesaplama değildir. Örneğin, yüksek rütbeli bir düşman lideri için kaç çocuğun feda edilebileceği bir zor orantılılık sorunudur. Bir makine mümkün olan en az sayıda ikincil hasarla en doğru saldırı türünü hesaplayabilir. Ancak, ilk etapta istenen hedefe herhangi bir ölümcül güç uygulanıp uygulanmayacağını hesaplayamaz. Bu hususlar göz önünde bulundurulduğunda, eleştirmenler OSS'lerin ne muharipleri muharip olmayanlardan ayırabileceğini ne de somut askeri avantaj açısından ne kadar ikincil zararın kabul edilebilir olduğunu hesaplayabileceklerini savunmaktadır. Sonuç olarak, bu sistemlerin "arkalarında bir yığın masum kurban" bırakabileceklerini belirtirler (Birnbacher, 2016, s. 118).

OSS karşıtları tarafından dile getirilen bu kaygılar, OSS'lerin beklenmedik koşullar, düşman davranışı, hava koşullarındaki değişiklikler gibi işlenmesi gereken çok sayıda girdinin bulunduğu düzensiz ve karmaşık ortamlarda nasıl çalışacağının öngörülememesinden kaynaklanmaktadır. Bu öngörülemezlik özelliği, OSS'lerin ayrım ve orantılılık ilkelerine uygunluğuna ilişkin eleştirileri de beraberinde getirmiştir.

3.1. İnsan'dan daha insan: OSS'lerin avantajları

Öte yandan savunucular, askerler ve mevcut askeri teknolojilerin de savaş alanında öngörülemezlikten mustarip olduğundan, OSS'lerin öngörülemezliğinin bu silah sistemlerini yasaklamak için meşru bir neden olamayacağını iddia etmektedir. Aslında, OSS'lerin savaştaki öngörülemezliği azaltacağını ve buna bağlı olarak savaş suçlarının, ikincil hasarların ve sivil kayıpların sayısını azaltabileceğini savunmaktadırlar.

Askeri robotik alanında çalışan önde gelen bir robotikçi ve roboetikçi olan Ronald Arkin'e göre bu makinelerin insanlara kıyasla daha etik olacaklardır. İlk olarak, makineler kendilerini korumakla ilgilenmezler ve gerektiğinde kendilerini kurban edebilirler. İkinci olarak, sensörleri insanların şu anda sahip olduğundan daha hassas gözlem yeteneği sağlamaktadır. Makinenin duyusal yetenekleriyle bağlantılı olarak, OSS'ler ölümcül eylemde bulunmadan önce çeşitli kaynaklardan gelen daha fazla bilgiyi hızla işleyebilir ve bir insanın gerçek zamanlı olarak yapabileceklerinden daha iyi performans sergileyebilir. Üçüncü olarak, OSS, kararlarını etkileyebilecek öfke ve intikam gibi duyguları olmadan programlanabilir. Son olarak, LAWS sadece insan askerlerden daha etik davranma potansiyeline sahip olmakla kalmaz, aynı zamanda savaş alanındaki etik davranışları bağımsız ve objektif olarak izleme ve ihlalleri rapor etme kabiliyetine de sahip olabilir, Arkin'e göre, "sadece bu özellik bile muhtemelen insanların etik ihlallerinde bir azalmaya yol açabilir" (2009, s. 29-30).

3.2. Öngörülemezlik

Öngörülebilirlik, bir makinenin eylemlerinin ne kadar tahmin edilebileceği anlamına gelir. Holland Michel (2020), otonom sistemlerde öngörülebilirliğin iki farklı yönü olduğunu savunmaktadır: "teknik" ve "operasyonel öngörülebilirlik" (2020, s. 5). Teknik öngörülebilirlik makinenin özelliğidir. Yani, sistemlerde kullanılan belirli programlama tekniklerine bağlıdır. Öte yandan, operasyonel öngörülebilirlik, otonom sistemlerin faaliyet göstereceği ortamın ve durumların karmaşıklığına atıfta bulunur.

Teknik öngörülebilirlik, belirli bir görevi otomatikleştirme teknikleri gibi sistemin teknolojik yeteneklerine dayanır. Örneğin, kural tabanlı algoritmalar, çevresel değişikliklere uyum sağlayan makine öğrenimi algoritmalarından daha öngörülebilir olabilir.

Sistem makine öğrenimi algoritmalarını kullandığında ortaya çıkan önemli bir problem vardır: açıklanabilirlik. Açıklanabilirlik, bu sistemlerin eylemlerinin altında yatan mantığın anlaşılmasındaki zorluk nedeniyle otonom sistemlerin kullanılmasında bir risk oluşturmaktadır. Basit bir ifadeyle açıklanabilirlik, bir sistemin neden belirli bir karar verdiğinin veya neden belirli bir şekilde hareket ettiğinin anlaşılamaması sorunudur. Bu sorun genellikle kara kutu sorunu olarak adlandırılır. Bu, insanların kararlardan tamamen habersiz olduğu anlamına gelmemektedir. Örneğin, yüz tanımada, programcılar sistemin yüzleri tanımayı öğrendiğini bilirler. Program iyi çalışıyor ve yüzleri güvenilir bir şekilde tanıyor olsa da, bu öngörülebilir olduğu anlamına gelmez. Kara kutu vakalarında, programcılar sistemin yüzleri nasıl tanıdığını ve programın yüz tanımak için hangi parametreleri veya özellikleri kullandığını bilmezler. Bu durumda, yüz tanıma sonucunun arkasındaki süreç açıklanabilir değildir. Nihayetinde bu sorun öngörülemezliğe neden olur çünkü kararlarının ardındaki süreç hakkında bilgi olmadan makinenin gelecekteki senaryolarda nasıl davranacağını tahmin etmek artık mümkün değildir. Ancak bu hususlar, otonom sistemlerde sistemin teknolojik özelliklerine bağlı olan bir tür öngörülemezlikle ilgilidir.

Teknolojik öngörülemezlikten bağımsız olarak, her tür otonom silah sistemi bir dereceye kadar "operasyonel öngörülemezlik" ortaya çıkarır (Holland Michel, 2020, s. 5). Operasyonel öngörülemezlik, operasyonel ortamın sorunudur. Örneğin, otonom silah sistemleri çok sayıda girdinin bulunduğu karmaşık ve dinamik ortamlarda çalışmak zorunda kalacaktır. Dost-düşman davranışları, diğer askeri robotlar, coğrafi varyasyonlar, hava koşulları ve bunların kombinasyonları gibi çok çeşitli girdiler göz önüne alındığında, makinenin tüm bu değişkenlerle nasıl etkileşime gireceğini öngörmek veya tahmin etmek ve makineyi tüm koşullar için önceden programlamak mümkün olmadığından, OSS'ler operasyonel anlamda doğası gereği öngörülemezdir. Dolayısıyla, operasyonel anlamda öngörülemezlik, hem kural tabanlı hem de makine öğrenmesi kullanan sistemler için kolayca üstesinden gelinemeyecek bir sorundur çünkü açıklanabilir yapay zeka'ya ulaşılsa bile, yani sistemin belirli bir kararı nasıl verdiğini veya belirli bir şekilde nasıl hareket ettiğini açıklamak mümkün olsa bile, otonom bir sistemin operasyonları sırasında karşılaşabileceği tüm potansiyel durumları tahmin etmenin zorluğu nedeniyle operasyonel öngörülemezlik devam etmektedir.

Bölüm 4: Ahlaki Sorumluluk

Her ne kadar OSS'leri savunanlar hali hazırda savaşta insanların da öngörülemez hareketler sergilediğini belirtip öngörülemezliğin OSS'leri yasaklamak için yeterli bir sebep olmadığını belirtse de insanların öngörülemezliğini otonom makinelerin öngörülemezliğinden ayıran önemli bir faktör vardır. İnsanlar suç işlediklerinde bu suçun sonuçlarına maruz kalırlar, yani suçlarının sorumluluğunu taşırlar. Otonom silahlar söz konusu olduğunda, sorumluluğun nerede olduğunu belirlemek zor olabilir ve hatta "sorumluluk boşluklarına" yol açabilir (Matthias, 2004, s. 177). Yani, bir makine tarafından gerçekleştirilen ahlaki olarak zararlı eylemde birilerine sorumluluk atamak zorlaşmakta hatta imkansızlaşmaktadır. Sorumluluğun hem jus in bello hem de UIH için temel varsayım olduğu düşünüldüğünde, OSS'lerin eylemlerinde sorumluluk ataması önemli hale gelmektedir.

4.1. Ahlaki sorumluluk ve OSS

OSS bağlamında, bu sistemlerin eylemlerinin arkasındaki aktörler üç katmana ayrılabilir. İlk katmanda, eylemlerin gerçekleştirilmesinde doğrudan rol oynayan makinenin kendisi yer almaktadır. İkinci katmanda makinenin kullanıcıları vardır. Bu, makineyi kullanmak için belirli niyetlerle karar veren bir insan komutan olabilir. Son olarak, üçüncü katman makinenin tasarımcılarıdır. Tasarımcıların niyetleri makinenin tasarımına, geliştirilmesine ve programlanmasında görülebilinir. Bu üç katman OSS'lerin eylemleri için potansiyel ahlaki sorumluluk konumları olarak tanımlanabilir. Bu üç katmana göre, robotun kendisi, robotu aktive eden komutan, ve tasarımcılar üzerinden ahlaki sorumluluk argümanı incelenebilir.

İlk olarak robot'un kendi hareketlerinden sorumlu olup olamayacağına bakabiliriz. Genel olarak ahlaki sorumluluk insana özgü olan özelliklere, örneğin bilinç, sahip olmayı gerektirir,. Sparrow'a göre de ahlaki sorumluluk acı çekme özelliğini gerektirir çünkü ahlaki sorumluluk cezalandırılabilirliği gerektirir ve birini cezalandırmak da o kişinin acı çekebildiğini gösterir. Sparrow'a göre makinalar acı çekemediği için hareketlerinin sonucunda cezalandırılamazlar. Bu nedenle de makinalar kendi hareketlerinden sorumlu tutulamazlar.

Komutanlar ve tasarımcılarda OSS'lerin eylemlerinden ötürü sorumlu tutulamazlar çünkü bu makinalar ne komutanlarının ne de tasarımcılarının tahmin edemeyeceği hareketlerde bulunacaklardır. Önceki bölümlerde tartışıldığı üzere, bir makineyi karmaşık görevler için programlamak belirsizliği beraberinde getirir, çünkü makine tasarımcıları ya da kullanıcıları tarafından öngörülmeyen veya amaçlanmayan davranışlar sergileyebilir. Makina öğrenmesi kullanıldığında bu belirsizlik daha da artar çünkü makine çevresinden öğrenir ve çevresine uyum sağlar. Bu nedenle, tasarımcıları ve komutanları kontrol edemedikleri eylemlerden sorumlu tutmak haksızlık olacaktır.

4.2. Vekaleten Sorumluluk

Ahlaki sorumluluk genellikle fail ile eylem arasında doğrudan bir ilişki gerektirir, yani fail ile eylem arasında hem zaman hem de mekan açısından asgari bir mesafe vardır. Ancak, birçok durumda, fail ile eylem arasında doğrudan bir bağlantı olmayabilir. Bu ahlaki sorumluluk anlayışı, bir failin başka bir varlığın eylemleri veya davranışları için sorumluluk taşıdığı dolaylı sorumluluktur. Bu başka bir varlık başka bir insan, insan olmayan bir hayvan, bir kolektif veya bir robot olabilir. Dolaylı sorumluluk, iki varlık arasında, Goetze'nin "ahlaki dolanıklık" olarak adlandırdığı özel bir ilişki türü nedeniyle ortaya çıkar (2021, s. 220).

Tasarımcılar otonom bir sistemin eylemlerini etkiler. Tasarımcıların niyetleri, bir anlamda, OSS'lerde mevcuttur. Örneğin Goetze, tasarımcıların niyetlerinin "[makine öğrenme sisteminin] eğitimi başarıyla tamamlandığında,... eğitim veri setinin seçilmesi, ödül fonksiyonunun oluşturulması, hiperparametrelerin ayarlanması vb." üzerindeki kontrollerinde belirgin hale geldiğini iddia etmektedir (2022, s. 9). Yazılıma ek olarak, tasarımcıların mühimmat türü (bomba, mermi, vb.) gibi donanım seçimleri de tasarımcının failliğinin makine ve eylemleriyle ahlaki olarak iç içe geçmesine katkıda bulunur. Bu seçimlere ek olarak, otonom sistem tasarımcılarının bu makinelerin ortaya çıkaracağı zararlı davranışları ele almaları ve düzeltmek için adımlar atmaları gerekmektedir. Bu gibi nedenler nedeniyle tasarımcı ve sistem arasında ahlaki bir dolanıklık oluşur.

Vekaleten sorumluluk kendi içinde bulanık bir sorumluluk durumudur. Ancak formel bir tanım yoluyla vekaleten sorumluluğu daha anlaşılabilir kılabiliriz. Bu amaçla, OSS bağlamında, aşağıdaki tanımı kullanarak tasarımcıların nasıl OSS'lerin eylemlerinden sorumlu olabileceklerini açıklayabiliriz.

Bu açıklama için varsayımsal bir vakayı ele alalım. Teslim olan bir askeri (gayrimeşru bir hedef) öldüren OSS vakasında,

-(i) 'teslim olan asker öldürülür', 'bazı hedefler öldürülür veya hiçbir hedef öldürülmez' kümesinde ahlaki olarak zararlı bir sonuçtur;

-(ii) OSS nedensel olarak 'teslim olan askerin öldürülmesine' katkıda bulunur;

-(iii) Tasarımcı gönüllü olarak OSS ile bir 'tasarım' ilişkisi içerisindedir;

-(iv) 'bazı hedefler öldürülür veya hiçbir hedef öldürülmez' kümesi 'tasarım' ilişkisinin kapsamına girer

Öyleyse, tasarımcı 'teslim olan askerin öldürülmesinden' dolaylı olarak sorumludur.

Bu durumda, (i) teslim olan askerin öldürülmesinin OSS'nin neden olabileceği potansiyel sonuçlar kümesinde olduğu anlamına gelir, yani bu sonuç 'bir hedef öldürülür veya hiçbir hedef öldürülmez' genel kümesi içerisindedir. (ii) OSS'nin olaya nedensel katkısını ifade eder. (iii) özellikle önemlidir, çünkü yukarıda tartışıldığı gibi, tasarımcının tasarım aşamasında sistemin yazılımı ve donanımı üzerindeki niyetleri ve seçimleri, OSS'nin savaşta nasıl çalışacağını tamamen belirlemese de etkiler. Bu aynı zamanda tasarımcının OSS'leri üreten şirkete katılımındaki niyetlerini de içerir. (iv) için, ahlaki açıdan ilgili ilişki, tasarımcı ile OSS arasındaki ‘tasarlama’ ilişkisidir. OSS üretiminin genel amacı hedefleri vurmaktır ve bu da OSS’nin çalışması sırasında bazı hedefleri vuracağını belirtir. Tasarımcının ne belirli bir sonuç üzerinde doğrudan kontrolü vardır ne de OSS’nin hangi tekil hedefe angaje olacağının farkında değildir. Ancak, tasarımcının ahlaki boyuttaki tasarım ilişkisi nedeniyle söz konusu sonuç üzerinde ahlaki anlamda hala kontrolü bulunmaktadır.

Sonuç olarak, bu hususlar tasarımcıların OSS’nin eylemlerinden ahlaki olarak sorumlu tutulabileceği yolları göstermektedir. Bu durum, bazı faillerin OSS’nin eylemleri üzerinde dolaylı da olsa kontrol sahibi olduğu önermesini kanıtlayarak sorumluluk boşluğu argümanını çürütmektedir. Dolayısıyla sorumluluk boşluğunun üstesinden gelinebilir.

5. Sonuç

Bu tezde, sorumluluk boşluğu sorununa bir çözüm olarak, sorumluluğun her zaman direk kontrolün varlığı koşuluna bağlı olarak atanmadığını öne sürülüyor. Ahlaki sorumluluk atamaları için doğrudan kontrolün her zaman gerekli olmadığını göstermek için, bir failin bir başkasının eyleminden sorumlu tutulabileceğini gösteren vekaleten sorumluluk kavramını öne sürdüm. Vekaleten sorumluluk belirsiz bir kavramdır çünkü bu iki fail arasında doğrudan bir bağlantı veya ilişki olmasa bile bir faili diğerinin eylemlerine bağlamayı amaçlar. Vekaleten sorumluluğun doğasında var olan belirsizliğin üstesinden gelmek için Glavanicova & Pascucci (2022) tarafından önerilen formel bir vekaleten sorumluluk tanımının değiştirilmiş bir versiyonunu kullandım ve bu tanımı OSS vakasına uyguladım. Bu tanımı izleyerek, OSS tasarımcılarının, yarattıkları tasarım ile olan ahlaki ilişkileri nedeniyle OSS’nin neden olduğu sonuçlardan ahlaki olarak sorumlu tutulabilecekleri sonucuna vardım. Dolayısıyla, OSS tasarımcılarının dolaylı sorumluluğu, OSS bağlamında sorumluluk boşluğu sorununun üstesinden gelmektedir.

B. THESIS PERMISSION FORM / TEZ İZİN FORMU

(Please fill out this form on computer. Double click on the boxes to fill them)

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences ☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ☒

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics ☐

Enformatik Enstitüsü / Graduate School of Informatics ☐

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences ☐

YAZARIN / AUTHOR

Soyadı / Surname : GÜLMEZ
Adı / Name : Salih
Bölümü / Department : Felsefe / Philosophy

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English): Military Robots: Ethics of Lethal Autonomous Weapon Systems

TEZİN TÜRÜ / DEGREE: **Yüksek Lisans** / Master ☒ **Doktora** / PhD ☐

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ☒
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two years**. * ☐
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. * ☐

***** *Enstitü Yönetim Kurulu kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir. / A copy of the decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.*

Yazarın imzası / Signature .... **Tarih** / Date ....
(Kütüphaneye teslim ettiğiniz tarih. Elle doldurulacaktır.)
(Library submission date. Please fill out by hand.)

Tezin son sayfasıdır. / This is the last page of the thesis/dissertation.