Deceptive AI machines on the battlefield: Do they challenge the rules of the Law of Armed Conflict on military deception?

Eleftherios Chelioudakis
Snr. 2006546 / Anr. 518189
Tilburg University

Abstract

Over the ages deception has been a traditional and crucial instrument for conducting warfare. However, military deception is not always acceptable. The Law of Armed Conflict distinguishes it in two categories, “Ruses of War” which refer to acts of deception that any commander can use to confuse the enemy regarding the military situation, and “Perfidy”, which refers in essence to outright treachery or the breach of good faith and is considered to be a war-crime. Machines of Artificial Intelligence which are capable of deceiving have been developed over the last years (2010-2014) by scientists. These deceptive AI machines, as we call them, have the ability to mislead opponents in a variety of ways and could be a desirable addition for military units. This research aims at examining whether there are possible frictions between the use of deceptive AI machines on the battlefield and the rules of the Law of Armed Conflict on military deception. By analyzing the deceptive capacities of these machines, the rules of the Law of Armed Conflict on military deception, and the problematic situations that arise the author concludes that the rules of the Law of Armed Conflict on military deception enjoy a high degree of flexibility and are not challenged by the deceptive AI machines.

Keywords: deceptive AI machines, robotic deception, Artificial Intelligence, military deception, Law of Armed Conflict, Law of War, International Humanitarian Law

Introduction

From ancient Chinese military treatises, such as "The Art of War" (Sun Tzu) to military manuals of the modern era, deception is considered to be critical and essential for conducting warfare (Handel, 2006). It has been a traditional instrument, which any commander aiming to lead the troops to victory would implement, in order to confuse and defeat the enemies. Misleading your opponent in a military situation, by hiding your actual location, the strength of your troops, your plans and intentions, are ways of using deception effectively on the battlefield. However, according to the Law of Armed Conflict (this is the term that will be used throughout the thesis, also known as "International Humanitarian Law", and the "Laws of War") deception in warfare is not always acceptable. The Law of Armed Conflict distinguishes two categories of military deception, "Perfidy" and "Ruses of War". The term "Ruses of War" refers to acts of deception that any commander can use to confuse the enemy regarding the military situation, while the term "Perfidy" refers in essence to outright treachery or the breach of good faith (The International Committee of the Red Cross, 2002).

Prominent computer scientists and roboticists around the world, such as R.C. Arkin, D. Floreano, S.A. Morin, K. Terada, and A.R. Wagner, have recently (2010-2014) developed machines of Artificial Intelligence, which are capable of deceiving. The term that will be used throughout this thesis referring to this type of technology is "deceptive AI machines". This, as well as the terms "Artificial Intelligence" and "robotic deception", will be analyzed in depth in later chapters. However, in order to provide a clear first description of this technology, deceptive AI machines can be thought of as systems that perceive the world around them, and based on this perception, can independently and unexpectedly perform a range of behaviors, such as deceiving humans or other machines.

The deceptive AI machines have the ability to mislead opponents in a variety of ways and could be a desirable addition for military units. As robots become more present in the future of the military, robotic deception can provide new advantages for the military force implementing it (US Department of Defense, 2009). However, these machines and their deceptive capabilities must be deployed in accordance with the Law of Armed Conflict and fthe rules It sets regarding the use of military deception. Outright treachery or the breach of good faith, are considered to be "Perfidy", and thus unacceptable forms of military deception. Acts of "Perfidy" qualify as war crimes (Rome Statue, 1998, art. 8b), and therefore it is crucial to ensure that these machines do not perform such acts. At first glance, some of the dilemmas that emerge are whether deceptive AI machines of the state of the art are allowed, by the Law of Armed Conflict, to participate in hostilities, whether their acts can be characterized as perfidious, how to make sure that they will not proceed to perfidious acts since they act unexpectedly, or who is to be held responsible if they proceed to a violation of a rule of the Law of Armed Conflict.

This thesis aims at examining whether there are possible frictions between the use of deceptive AI machines and the rules of the Law of Armed Conflict on military deception, by firstly analyzing the deceptive capacities of these machines, secondly, the rules of the Law of Armed Conflict on military deception, and thirdly, the problematic situations that arise. The subject matter of this assessment is whether the use of the deceptive AI machines challenges the rules of the Law of Armed Conflict on military deception. Therefore, the author will examine only the elements that have a direct relevance for this research, i.e. excluding from the assessment other situations that can emerge from the use of Artificial Intelligence on the battlefield. For instance, the killing, wounding, or capturing of a combatant without the use of a deceptive tactic, or murdering of a civilian. Moreover, possible suggestions for revising the

existing Law of Armed Conflict are also outside of this dissertation's scope. Therefore, the goal of this thesis is to answer the research question:

"Does the use of deceptive AI machines in warfare challenge the rules of the Law of Armed Conflict on military deception, and if it does what are these challenges?"

The research question derives from the existing inadequacies of the current literature, where no sources examining the possible frictions between the use of deceptive AI machines and the Law of Armed Conflict could be detected.

There is detailed literature and a considerable volume of case-law and legislation on the field of military deception. Notable examples of such literature are: the works of authors like J. M. Mattox (1998), or D. Jackson and K. Fraser (2012), research reports of the International Committee of the Red Cross (2002), treaty law and customary law like "The Hague Conventions" (1899 & 1907), "The Additional Protocol I to the Geneva Convention of 1949" (1977), and "The List of Customary Rules of International Humanitarian Law" (2005), and caselaw such as the case "Nikolić, Momir IT-02-60/1" (International Criminal Tribunal for the former Yugoslavia or ICTY, 2002).

Moreover, significant body of literature exists in the fields of Artificial Intelligence and robotic deception. For instance, concerning the field of Artificial Intelligence, works of J. McCarthy (1959), N.J. Nilsson (2010), S. Russell & P. Norvig (2016) and M. Cummings (2017), while in the field of robotic deception research studies of D. Floreano & L. Keller (2010), A.R. Wagner & R.C. Arkin (2011), K. Terada & A. Ito (2011), and J. Shim & R.C. Arkin (2012 and 2013). Additionally, the "Association for the Advancement of Artificial Intelligence" (AAAI, formerly the "American Association for Artificial Intelligence") has offered over the last years (2011-2016) through the AAAI Fall Symposium Series a great number of research studies on deceptive AI machines. Notable examples are the works of S.

Fahlman (2015) and M. Clark and D. Atkinson (2013). However, the legal and technical accounts on Artificial Intelligence and military deception seem to never overlap.

Thus, as separate topics, there is detailed literature of military deception and of deceptive AI machines. However, a lacuna is identified in the existing body of knowledge. There does not exist a multidisciplinary research examining the technology of the deceptive AI machines under the legal framework of the Law of Armed Conflict and Its rules on military deception.

In the existing literature exists a somewhat relevant work which examines the legality of the use of non-autonomous machines on the battlefield (such as unmanned aerial systems, broadly known as "drones"), like the essays of C. Jenks (2010), or M. Schmitt (2011), and work which addresses the use of "Autonomous Weapons Systems", also known as "killer robots" in warfare. Essays of B. N. Kastan, (2013), W. Marra & S. McNeil (2013), E. Lieblich & E. Benvenisti (2014), and R. Crootof (2014 and 2015) examine the legality of these machines and wonder how their use could possibly alter the distribution of constitutional war power. Nevertheless, again none of these examine whether there is a conflict between the use of deceptive AI machines on the battlefield and the rules of the Law of Armed Conflict on military deception.

To answer the main research question, the author will divide it into three sub-questions:

1. "What is a deceptive AI machine?",
2. "What is permissible and what is forbidden regarding the use of deception in warfare under the rules of the Law of Armed Conflict", and
3. "What are the problems that arise from the use of deceptive AI machines on the battlefield?".

Each of these sub-questions will be the focal point for a chapter of this thesis, and by providing an answer to each of them the author will deliver a final answer to the main question.

Particularly, in the first chapter the author will provide a thorough assessment of the notions of "Artificial Intelligence" and "robotic deception", in order to offer to the reader an explanation of the two and a working definition of "deceptive AI machines". In the second chapter, the aim will be to deliver a comprehensive analysis of what is permissible and what is forbidden regarding the use of deception in warfare. Subsequently, the author will analyze the notions of "jus ad bellum" and "jus in bello", and offer a detailed inquiry about the differences between the two aspects of military deception, "Perfidy" and "Rouses of War". Finally, in the third chapter of the essay the author will offer a deeper analysis of the rules of the Law of Armed Conflict as applied to the deceptive AI machines that were examined in the first chapter. The problems that arise from the use of deceptive AI machines will be examined, and the author will assess whether these problems can lead to challenges for the rules of the Law of Armed Conflict on military deception. At the end of this thesis, the author will answer whether the use of deceptive AI machines challenge or do not challenge the rules of the Law of Armed Conflict on military deception.

Throughout the thesis, the methodology that will be used is the Doctrinal (Black Letter) legal analysis, while the author will focus his research on the Law of Armed Conflict and specifically Its rules on the use of military deception. This is because the Law of Armed Conflict, as part of the Public International Law, is composed by treaties and customary law that address the appropriate means and methods for conducting warfare. Thus, the Law of Armed Conflict is one of the most important sources of state's obligations regarding the conduct of warfare in general, and the use of military deceptive tactics on the battlefield in particular. The author will exclude from his analysis other treaties and customary rules of the

Law of Armed Conflict that do not have a direct relevance for this research, such as provisions of treaties protecting cultural property and civilians etc.

In the first chapter, sources such as leading textbooks and scientific research studies in the fields of Artificial Intelligence and robotic deception will be taken into consideration. Also, academic articles and essays of computer scientists, legal scholars and scholars of cognitive sciences will be used in order to help the reader become familiar with the terminology and acquire clarity regarding the subject. Examples of the above are the textbook "Artificial Intelligence: A Modern Approach" by S. Russell & P. Norvig (2016), studies on Artificial Intelligence of N.J. Nilsson (2010), M. Cummings (2017), or J. Zittrain, J. (2017), and researches on deceptive AI machines of D. Floreano & L. Keller (2010) or J. Shim & R.C. Arkin (2012 and 2013).

In the second chapter, the research will be based both on primary and secondary sources of the Law of Armed Conflict. Primary sources that will be examined are international treaties, rules of customary international humanitarian law, and court decisions. For instance, "The Hague Conventions" (1899 & 1907), "The Additional Protocol I to the Geneva Convention of 1949" (1977), Customary Rules of the Law of Armed Conflict on "Ruses of War" and "Perfidy", and cases such as the one of "Nikolić, Momir IT-02-60/1" (ICTY, 2002). Secondary sources that will be used are works that interpret and analyze the above mentioned primary sources, such as the work of J. M. Mattox (1998) or research reports of the International Committee of the Red Cross (2002) and documents of historical value such as the "The Brussels Declaration of 1874". This to present to the reader with the elements that determine whether a deceptive tactic is acceptable or not.

Finally, in the third chapter the author will offer a deeper analysis of the rules of the Law of Armed Conflict as applied to the deceptive AI machines which were examined under

the first chapter. Examples of primary sources that will be examined are the "Additional Protocol (I) to the Geneva Conventions of 12 August 1949" (1977) and Its provisions on the development of new means or methods of warfare (art. 36), and Customary Rules of the Law of Armed Conflict on the "combatant status". Examples of secondary sources in this chapter are the works of Detter de Lupis Frankopan (2013), as well as reports of the Advisory Service of the International Committee of the Red Cross on International Humanitarian Law (2004), and the

A detailed analysis of the specific sources that will be used (primary and secondary) will be placed by the author in the beginning of chapter two and chapter three (please see p. 23 – 24 and p. 36 – 37 respectively).

Chapter 1: An explanation of the notion of the deceptive AI machine

In the first chapter of this thesis the attention will be directed to the sub-question "What is a deceptive AI machine?". In order to offer a sufficient answer to this query, a thorough assessment of the notions of "Artificial Intelligence" and "robotic deception" will take place. The goal in this chapter is to place an explanation of these two notions and a working definition of the "deceptive AI machines" at the reader's disposal. By doing this, the reader will become familiar with the terminology and acquire clarity regarding the subject matter of this thesis.

1.1 What is Artificial Intelligence? One question - a lot of answers

Films and novels during the 20th and 21st century introduced the wider public to the notion of Artificial Intelligence. The transition of this notion from cinemas screens and novel pages to real-life, has long being awaited. Now, in 2017, a reality is witnessed in which humans start to coexist with machines of Artificial Intelligence and interact with them on a daily basis. Digital assistants are an example of such interaction. Individuals around the

world use Microsoft's Cortana, Amazon's Alexa, or Google's Home Assistant to find answers to a wide variety of questions and concerns. Additionally, machines of Artificial Intelligence are attempting to enter other industries, such as transportation, healthcare, and defense. Without human intervention, autonomous vehicles will drive us to our destinations, smart algorithms will fight diseases, and autonomous weapons will enter the battlefield. But what exactly is a machine of Artificial Intelligence and what is it capable of doing? Is there a universally acceptable definition that can be used or not? In this first part of chapter one, the author will offer answers to such questions and bring clarity to the notion of Artificial Intelligence.

To begin with, Artificial Intelligence is a complex notion. Thus, legal scholars, cognitive scientist and engineers, have great difficulties in reaching a consensus on a specific and commonly accepted definition. In this respect, it is noted that there is lack of coherence regarding the definition of Artificial Intelligence even among professionals of the same expertise. Additionally, the term "Artificial Intelligence" is not the only one widely used to describe this type of technology. On the contrary, there are other terms describing the same technology that are widely accepted, such as "Deep Learning", "Machine Learning", "Autonomous Systems", "Artificial Consciousness", "Computational Cognition" among others. For reasons of consistency, the author will only use the term "Artificial Intelligence" throughout. In the subsequent paragraphs, some of the existing thoughts on what Artificial Intelligence is will be provided, and by finding the common grounds in these thoughts, a working definition of Artificial Intelligence will be coined.

The assessment will start by examining the work of McCarthy, a prominent computer and cognitive scientist who is considered to be the father of Artificial Intelligence. Together with Minsky, Shannon and Rochester, they used for the first time the term Artificial Intelligence, in their written proposal for the Dartmouth summer research project (Rajaraman,

2014). According to his work, Artificial Intelligence is described as the science of creating machines or computer programs that are intelligent, while intelligence has the meaning that these machines or computer programs are capable of achieving goals in the real world (McCarthy, 1959). Therefore, the conceptual core of their definition is that intelligence is inevitably bound with efficacy, which is the capacity to realize a purpose or to achieve a goal, or the ability to solve a problem. However, another computer scientist, Nilsson, offers a different standpoint for a machine's intelligence. Based on his definition, Artificial Intelligence is the creation of intelligent machines, but intelligence is the "quality that enables an entity to function appropriately and with foresight in its environment" (Nilsson, 2010, p.13). Thus, by introducing the notion of foresight, Nilsson takes a step further as to characterize a machine as an intelligent one. He demands that it not only is functional but also acts with prudence, be cautious and have the capacity to reflect on the impacts of its actions.

In "Artificial Intelligence: A Modern Approach", Norvig and Russell refer to Artificial Intelligence as "agents that receive percepts from the environment and perform actions". This definition might seem to be quite simple and generic. However, its importance lies exactly in these two characteristics, since in this manner it manages to enclose under the term Artificial Intelligence agents that base their reactions both on real-time planning, and/or a decision-theoretic system. (Norvig and Russell, 2016, Preface, p.8).

Norvig and Russell also provide for a thorough analysis of the existing definitions of Artificial Intelligence. These definitions are divided into four categories based on the manner in which Artificial Intelligence is approached. The first category includes definitions that focus mainly on the aspect of thinking like a human. An example is Bellman's definition, according to which machines are considered to be intelligent when their activities are associated with human thinking, e.g. learning, problem-solving or decision-making

(Bellman,1978). In the second category definitions that have as their focal point the machine's capacity of thinking rationally, are found. For instance, according to Winston, machines of Artificial Intelligence are those that are able to perceive, reason, and act (Winston, 1992). The third category comprehends definitions that are centered around machines acting like a human. A definition that epitomizes this category is the one of Rich and Knight, which states Artificial Intelligence could be accomplished when computers are able to do things that humans are currently better at doing (Rich and Knight, 1991). Lastly, the fourth category encompasses the definitions that have as their conceptual core a machine's capability of acting rationally. A good example of a definition in this category is Poole's, which is based on the idea that Artificial Intelligence is about studying the design of intelligence agents (Poole et al., 1998).

A different approach of Artificial Intelligence is presented by Boden. In her work, she defines Artificial Intelligence as the use of programming techniques and computer programs in order to cast light on the principle of human thought and intelligence (Boden, 1987). Therefore, she suggests that through examining Artificial Intelligence, scientists can understand human intelligence and rationality further, by obtaining a better understanding of the way humans perceive their environment, process information obtained, and make decisions.

Additionally, the think-tank Transpolitica explains the notion of Artificial Intelligence as:

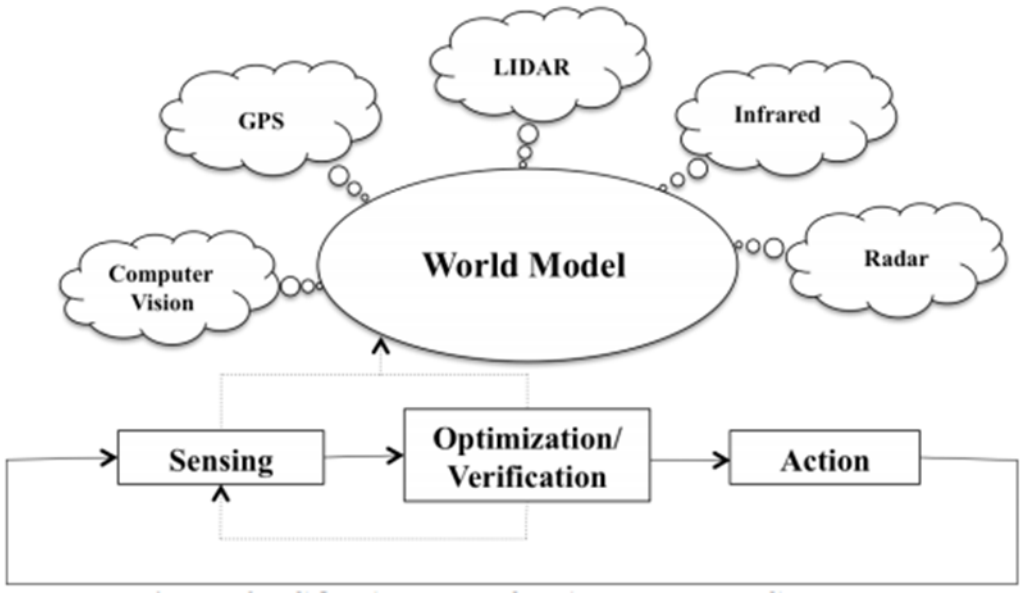
"a set of statistical tools and algorithms that combine to form, in part, intelligent software that specializes in a single area or task. This type of software is an evolving assemblage of technologies that enable computers to simulate elements of human behavior such as learning,

reasoning, and classification" (House of Commons of the United Kingdom Science and Technology Committee Report, 2016, p.6).

This definition has as its focal point the concept of "specialization". In the same way that humans become experts in various kinds of professions and sciences, intelligent software programs need to be able to have deep expertise and increased specialization as well. In order to develop such a capacity, these software programs need to process the data from their environment in a syllogistic order (reasoning), analyze and categorize them (classification), and improve through this procedure their performance (learning). An example of such an agent of Artificial Intelligent is AlphaGo, which masters the board game "Go".

On the contrary, a broad definition of Artificial Intelligence is provided by Zittrain. He acknowledges that definitions are to a certain extent labels, and therefore suggests that another suitable label for this technology is "just forms of systems that evolve under their own rules in ways that might be unexpected even to the creator of those systems, that will be used in some way to substitute for human agency" (Zittrain, 2017, p.1). In this definition, the concept of "unexpected" is mentioned. These systems do not have standard behaviors. They might have their own dependencies but at the same time they interact with people, other systems, and generally their environment, forming behaviors that are not always anticipated. Moreover, by using plain language and avoiding technical terms, he encompasses under the term of Artificial Intelligence, a variety of agents that base their functionality in different types of existing techniques such as "genetic algorithms", "reinforcement learning", and "neural networks". For a quick understanding of these different techniques, "genetic algorithms do not allow agents to learn during their lifetimes, while neural networks allow agents to learn only during their lifetimes. Reinforcement learning allows agents to learn during their lifetimes and share knowledge with other agents" (Patel, 2017, p.1)

Furthermore, another approach to the concept of "unexpected" is provided by Cummings. She together with Hutchins, Draper and Hughes created a model for determining the self-confidence of agents of Artificial Intelligence. According to this model (called "Sensing-Optimization/Verification-Action" model- "SOVA"), there are three stages for a machine of Artificial Intelligence to perform an assigned task: the "sensing" stage, in which the machine creates a perception about the world that surrounds it through series of sensors, the "optimization/verification stage" in which the machine's algorithms determine what actions the machine should take in order to complete its assigned task with safety and efficiency, and "action" stage in which the machine based on its preprogrammed scripts determine the execution of the task (Hutchins, Cummings, Draper, & Hughes, 2015).



[1]

Based on the model depicted above, Cummings explains that an automated system, which in not a machine of Artificial Intelligence, always delivers the same output for each

[1] SOVA model, Source: Hutchins, Cummings, Draper and Hughes 'Representing Autonomous Systems' Self-Confidence Through Competency Boundaries' (2015). "LIDAR" stands for "Light Detection and Ranging"

input of data, except in cases where something fails during the procedure. On the contrary, a machine of Artificial Intelligence even given the same input of data, "will not necessarily produce the exact same behavior every time; rather, will produce a range of behaviors" (Cummings, 2017, p.4). Consequently, it is understood that a machine of Artificial Intelligence is not constrained by its creator, but instead it enjoys, at a greater or lesser degree, independent behavior.

In light of the assessment of the notion of Artificial Intelligence and the different definitions that have been examined, it can be concluded that there are a lot of different or recurrent elements in these definitions. These machines receive information from the environment and create a perception about the world that surrounds them (element of perception), and they perform actions in order to achieve goals in the real world (element of performance). Additionally, a variety of focal points can be seen in these machines, such as their characteristic to evolve under their own rules (element of independency), and to act in ways that might be unexpected even to their creator (element of unexpectedness). Also, it is understood that these machines could specialize in a variety of different areas or tasks and produce a range of behaviors (element of a wide range of behaviors), and that the way they behave simulates elements of intelligence such as learning, reasoning, and classification (element of intelligence).

For the purposes of this thesis, and based on the ideas that have been examined in this first part of chapter one, the machines of Artificial Intelligence will be defined as "systems, which perceive the world around them and based on this perception are able to perform, in an independent and unexpected way, a range of behaviors that naturally require intelligence". The next step in this assessment, is to analyze the notion of robotic deception. Subsequently, the author will answer the sub-question of this chapter, and create a working definition for the "deceptive AI machines".

1.2 Robotic Deception: When robots lie

In this second part of chapter one, the notion of "robotic deception" will be examined. However, first the author will define what is considered to be deception in a human – human interaction. As was examined in the previous part of this chapter, machines of Artificial Intelligence are considered to act and/or think like humans, or to act and/or think in a rational way. Therefore, the understanding of when a human is indeed deluded, and how the human intelligent could be used in order to deceive others, is a crucial step towards explaining the notion of deceptive AI machines.

According to Lynch, "deception is a misleading that is willful or non-accidental" (Lynch, 2009, p.191). Similarly, Carson suggests that:

"deception requires some kind of intention to cause others to have false beliefs. A person *S* deceives another person *S1* if, and only if, *S* intentionally causes *S1* to believe *x*, where *x* is false and *S* does not believe that *x* is true" (Carson, 2009, p.179).

Therefore, it is understood that the first important aspect of deception is the "intention". Someone needs to be willing to give another person a false belief about something. This can be done either by telling lies or by withholding the truth, since deception is often best accomplished by what is left unsaid. Additionally, Carson states that deception connotes success meaning deception is something that is believed. Thus, it is understood that the second aspect of deception is the "adoption of the fake belief". A person is deluded only if he/she truly believe that something false is true, and not if he/she pretends to do so or is not bothered whether something is true or false. Consequently, trust is the converse of deception and at the same time "a precursor for deception" (Arkin, 2011, p.1). Deception has also been defined as "a false communication that tends to benefit the communicator" (Bond, & Robinson, 1988, p.295). In this context, the third aspect of deception is "the benefit of the

liar". Deceptive behavior needs to be beneficial for the one who uses it and to result in an advantage for the latter.

Regarding robotic deception, firstly the research of Floreano and Keller (2010) will be examined. This work is not focusing on the robot-human relationship, but merely on a robot-robot interaction, since its focal point is the fact that deceptive capabilities could be developed in a group of homogenous artificial intelligent robotic agents, through the use of evolutionary robotics. The latter computational methodology "applies the selection, variation, and heredity principles of natural evolution to the design of robots with embodied intelligence. It can be considered as a subfield of robotics that aims to create more robust and adaptive robots" (Doncieux, Bredeche, Mouret, & Eiben, 2015, p.1). Specifically, these scientists conducted an experiment, in which generations of robots were competing for iconic food, the more time they spent in the food area, the more points they gathered. Each of these robots had one light emitter adjusted. When robots were near the food area, they emitted light, which attracted other robots to the area. Since robots were competing for the food and they did not want the food area to be crowded, the new generations of robots quickly selected to stop emitting light when they were close to the food area. In this manner, they concealed this information from the other robots and gained more points for themselves. This experiment demonstrated how deceptive behavior can emerge from simple rules (Floreano and Keller, 2010).

A different approach to robotic deception can be seen in the "camouflage robots", which were developed by scientists in Harvard University (Morin, Shepherd, Kwok, Stokes, Nemiroski, & Whitesides, 2012). Inspired by nature's soft living organisms, such as the squid and the octopus, these soft "camouflage robots" have the capacity to change their appearance and body color in order to match their environment or signal each other. Although the research of Morin et al. was focused on soft robots (machines fabricated from soft polymers

and flexible reinforcing sheets), their system could also interface with hard robots. Camouflage is a famous deceptive tactic used not only by various animals, but also by the military (Arkin and Shim, 2013).

Another research, this time focusing in creating a framework about how a machine of Artificial Intelligence could make decisions about "when to deceive" was conducted by the researchers Wagner and Arkin (2011). They implemented two theories, namely game theory and interdependence theory. This was to examine robotic deception and create an algorithm, which would enable a machine of Artificial Intelligence to decide when the use of deception was justified. The researchers focused on the actions, beliefs and communications of a robot (the deceiver and/or hiding robot) attempting to hide from another robot (the seeker and/or the deceived). Their first step was to teach the deceiver how to recognize a situation that warranted the use of deception. Such a situation had to satisfy two key conditions to warrant deception. First, there had to be a conflict between the deceiver and the seeker, and second, the deceiver had to benefit from the deception. Once a situation was deemed to warrant deception, the deceiver carried out a deceptive act by providing false communication to benefit itself.

To test their theory, the researchers created an experiment with two robots that acted autonomously. The deceiver was trying to hide, and the other robot (the seeker) was trying to find the hiding robot. In this experiment, there were three possible paths that the hiding robot could take, and each of these paths had in its beginning a marker. The hiding robot's tactic was to deliberately knock down the marker of a path and then change course and follow another path. Because of this, the seeker believed that the absence of a standing marker indicated that the hiding robot had taken that specific path. This false communication indicated a false hiding position. As a result, the seeking robot was deceived by the behavior of the hiding robot 75% of the time the experiment was conducted. The experiment failed

only when the hiding robot was incapable of knocking over the marker (Wagner and Arkin 2011). The research was funded by the "Office of Naval Research".

Additionally, Shim and Arkin (2012) conducted a scientific research, which was examining the way that a robot could make decisions on "how to deceive". In this research, the scientists were inspired by the deceptive behaviors of animals, specifically tree squirrels. These animals hide their food, often revisiting the locations for reassurance that it is still there. However, since squirrels have competitors for the food, they often use deceptive behavior in an attempt to protect their food stocks. When tree squirrels realize that they are being followed by other food competitors, they purposely visit locations that they have not stored their food. Thus, they deceive their competitors and drag them away from the locations where the food is actually stored. Moreover, if squirrels notice that some of the food is missing and consequently their hiding place is exposed, they carry the food away and store it to a new location.

The researchers managed to model these behaviors in small "squirrel- robots" and taught them to hide and protect virtual food. During this experiment, competitor robots were introduced with their own aim of finding the food. When the "squirrel-robots" noticed their competitors, they started to visit locations where their food was not hidden. In this way, the competitor robots were deceived as to the location of the food and were unable to find and claim it. This research was supported in part by the "Office of Naval Research" (Shim and Arkin, 2012).

Furthermore, Terada and Ito (2011) illustrated an experiment, where the robot partner was able to deceive the human one, by acting in a way that was in contrary with the expectations of the latter. Specifically, they conducted a test in which the human partner played a Japanese game for kids with the robot partner, "Darumasan ga Koronda" (called also

"Statues", or "Red Light-Green Light" in other countries). In this game, the robot faces away from the human partner towards the wall and so is unable to see the human. Then it says a special sentence whilst the human partner comes as close as possible to the robot partner while the sentence is spoken aloud. When the robot finishes this sentence, it turns around. The human partner then stops moving, and the robot partner checks the human one. If the human partner is not moving, the robot turns around again, and the whole process is repeated. This happens until the robot partner catches the human partner moving, or conversely the human partner manages to reach the robot. While the human partner is far away the robot says the special sentence in 5,5 seconds. But as the human partner comes closer and closer the robot partner, changes its behavior and there is a point when it says the special sentence in just 1 second. By doing this, the human partners were caught on the move by the robot, since they thought that they had more time to move towards the robot before the robot turned to face them. They did not expect the robot to say the special sentence that faster, because the latter had concealed its capability to say the special sentence in less than 5,5 seconds. All the human partners lost the game after this behavior was implemented by the robot and formed the impression that they had been successfully deceived (Terada and Ito, 2011).

The research studies on robotic deception that were examined in this second part of chapter one, show how scientists have created a breeding ground for deceptive AI machines. In a robot-human or a robot-robot relation an overall analysis of this deceptive capacity can be seen as the following: the deceiver robot partner will first gather data about its environment and the other partner, and form a belief about the current situation. Then as a second step, it will recognize whether in this situation the use of deception could be beneficial for itself. If this is the case, the deceiver robot partner will use verbal or non-verbal communication to make the other partner form an erroneous belief about the potential actions

and perceptions of the deceiver robot partner. Finally, the deceived partner is misled and the deceiver partner benefits from the resulting situation.

1.3 From machines of Artificial Intelligence to deceptive AI machines that are used on the battlefield

Through the assessments that were made in the first and second part of this chapter "Artificial Intelligence" and "robotic deception" were analysed. In this final part of chapter one, the final goal to coin a working definition for the "deceptive AI machines" will be achieved, based on the conclusions drawn from each of the two previous sub-chapters. Thus, deceptive AI machines will be defined as:

"systems which can perceive the world around them and based on this perception are able to perform, in an independent and unexpected way, deceptive behaviors that naturally require intelligence, such as making decisions regarding both when to deceive and how to deceive humans or other machines".

A reasonable question that arises, relates to the connection between the research studies that were examined regarding deceptive AI machines and warfare. A first answer to that query is that the scientists that conducted these studies admit themselves that one of the most appropriate uses of these machines is in military applications, where the use of deception is not only "laudable but also honorable" (Arkin, 2011, p.2). Specifically, based on their work, deceptive behaviors can be seen in various situations "ranging from warfare to everyday life" (Shim and Arkin, 2012, p1.). While in an interview by Arkin, these deceptive AI machines could be used in warfare because when an enemy is around:

" the robot could change where and when it patrols in an attempt to deceive his enemy, that can be another intelligent machine or a human. By doing as such, the robot could buy some time until reinforcements will finally come and assist it" (Ackerman, 2012, p.1).

Additionally, a second reason that connects the deceptive AI machines with the battlefield, is the fact that two of the examined studies, namely the ones from Wagner and Arkin (2011) and Shim and Arkin (2012), were funded fully and partly respectively by the "Office of Naval Research (ONR)". ONR is an organization within the "United States Department of the Navy" which promotes and organizes the science and technology programs of the U.S. Navy and Marine Corps. According to the "Naval Science and Technology Strategy", the investment priorities of this organization are "reflected in the allocation of funds across four components of ONR's strategic portfolio, and further aligned by mapping capability gaps to nine science and technology focus areas" ("Naval Science & Technology Strategy- Office of Naval Research", 2017, p.1). One of these focus areas is "Autonomy and Unmanned Systems", and thus, it is understood that since these research studies were funded by ONR, there is a high interest by the U.S. Navy and Marine Corps in deceptive AI machines.

In conclusion, it is understood that deceptive AI machines could be a desirable addition for some military units. However, these machines and their capacities must be in accordance with the Law of Armed Conflict, and the rules that the latter sets regarding the use of deception in military operations. In the following chapter, the goal will be to assess and understand these rules.

Chapter 2: An examination of when military deception is acceptable and when is not according to the rules of the Law of Armed Conflict

The aim of this chapter, is to offer the answer to the sub-question "What is permissible and what is forbidden regarding the use of deception in warfare under the rules of the Law of Armed Conflict?". The previous chapter, referred to the suggestions made by the researches of robotic deception, that one of the most appropriate uses of this technology is in

military applications, where deception is widely accepted, worthy of praise, and honorable. But, is this indeed always the reality? When is military deception acceptable according to the Law of Armed Conflict? To answer these questions, the research will be based on both primary and secondary sources of the Law of Armed Conflict.

As regards primary sources, in the subsequent paragraphs will be examined treaties that address the appropriate means and methods for conducting warfare. Specifically, "The Hague Conventions of 1899 and 1907" respecting the Laws and Customs of War on Land" and Their provisions on "Ruses of War" and "Perfidy" (art.22, 23, and 24), and the "Additional Protocol (I) to the Geneva Conventions of 12 August 1949, relating to the protection of victims of international armed conflicts" (1977) and Its provisions on "Ruses of War" and "Perfidy" (art. 37). Also, another treaty that will be analyzed is the Rome Statute of the International Criminal Court (1998), and Its provision qualifying "Perfidy" as a war crime (art. 8b). Additionally, customary humanitarian law will be examined, i.e. the Customary Rules on "Ruses of War" and "Perfidy" (Rule 57 and Rule 65, respectively of the 161 Rules of customary international humanitarian law[2]). Moreover, caselaw of ad hoc tribunals, such as the "International Criminal Tribunal for the Former Yugoslavia (ICTY, case "Nikolić, Momir IT-02-60/1, 2002), and of domestic courts that enforce and interpret the Law of Armed Conflict, such as the "General Military Government Court of the US Zone of Germany" (case United States v. Skorzeny, 1949) will be assessed.

Concerning secondary sources, in this chapter will be used works that interpret and analyze the aspects of the Law of Armed Conflict, as well as the above mentioned primary sources. For instance, the texts of Mattox (1998), Donovan, (2002), Rohde (2012), Jackson

[2] The numbering of the Rules is according with the most authoritative and wide-ranging compilation of customary International Humanitarian Law, i.e. Jean-Marie Henckaerts' and Louise Doswald-Beck's, Customary International Humanitarian Law (Cambridge University Press, 2005), Vol. 1, Rules and Vol 2, Practice.

and Fraser (2012), Detter de Lupis Frankopan (2013), Watts (2014), or Greer (2015), research reports of the Human Rights Watch (1995) and of the International Committee of the Red Cross (2002) and documents of historical value such as "The Lieber Code" (1863) and "The Brussels Declaration (1874)".

The aspects of the Law of Armed Conflict, and more specifically the notions of "jus ad bellum" and "jus in bello", will be examined in the first part of this analysis (2.2). In the second part of the analysis, there will be a thorough inquiry about the differences between the two different types of military deception, namely "Perfidy" and "Rouses of war" based on the provisions of customary and treaty law (2.3). In the third part, court cases on military deception will be analyzed (2.4), while in the final part will be stated remarks and conclusions of the whole assessment (2.5).

Before the start of the analysis, it will be interesting to have a look at one of the very first registered examples of military deception on the global stage, and see what can we learn from this incident (2.1).

2.1 Beware of Greeks bearing gifts[3]

Since ancient times, the use of deception in warfare has been a key element for the accomplishment of victories. Homer's "Iliad", the epic poem of the Trojan Wars, includes a notable example of the deceptive tactic of the Greek army, known as "the Trojan Horse". After a siege period of ten years, the Greek army unable to breach the strong walls of the city of Troy, found another way to defeat its enemies thanks to an idea of Odysseus. A wooden horse was built and the best soldiers of the Greek army were hidden inside. Outside of the horse was carved the phrase "For their return home, Greeks dedicate this as an offering to Athena" (Homer, n.d., Epit..5.16). During the night, Greeks left the horse, as a gift, outside of

[3] Virgil (29-19 B.C.), Aeneid (II, 49), In Latin: Timeo Danaos et dona ferentes

the walls of Troy and pretended to have gone back to Greece, when in reality they were hiding the rest of their army on a nearby island. The next day, the Trojans saw the wooden horse, which according to their tradition was a holy animal, and deceived by the carved sign, brought the horse inside their walls and celebrated their "victory". After the celebrations, and while the Trojans were sleeping, the Greek soldiers came out of the horse, killed the guards, opened the gates and the Greek army marched inside killing soldiers and citizens of Troy. The Greeks won this war, but the Gods never forgot their sacrilegious and heinous act. They cast their wrath upon all the main actors who played a role in this deceptive strategy, and a typical example was the deceitful Odysseus. Homer, in his second epic poem called "Odyssey" describes the trials and tribulations that Odysseus needed to go through after the end of the war, for a period of ten years, because of the use of this deceptive tactic, before his final return to his homeland (Homer, n.d.).

What can be seen in the myth described above, is the Greek army had gone beyond the boundaries. They pretended to retire from the war and offered a religious gift as a recognition of victory to their enemy. Hours later they killed the Trojans in their sleep. Lying in such a way in order to kill your enemy, was judged unacceptable by the wrath of the Gods. In current times, it is not the will of Gods, but the Law of Armed Conflict that define what is acceptable and was it is not regarding military deception. It will be interesting to try and find out whether today's society has changed the rules governing military deception or the same conclusion that was expressed by Homer in the ancient times is seen today.

2.2 The legal concepts of "jus ad bellum" and "jus in bello"

The first of the two aspects that comprise the landscape of the Law of Armed Conflict is the notion of "jus ad bellum". According to Jackson, philosopher of anthropology, "jus ad bellum", or in English the "right to (wage) war" "refers to a set of moral constraints of the

justifiability of resorting to particular military campaigns" (Jackson and Fraser, 2012, p.581). Therefore, it is understood that "jus ad bellum" answers the question of why to wage a war, and it specifies the requested criteria that justify a state's military action against another state. It sets the rules and the standards of when resorting to warfare is permissible and acceptable.

More specifically, Jackson describes six standards (Jackson and Fraser, 2012, p.581) that need to be met for a situation that allows a war to be just. First of all, there is a need of a "Just Cause". This is the premier element of the notion of "jus ad bellum" and means that the state's argument to wage a war shall be morally justifiable. An argument that qualify as such is the defense of innocent against an armed attack. The second standard is the "Right Intention" which draws its attention to the internal motivation of those who will engage in an armed conflict, which itself should be just. Thirdly, for a just war, there is the necessity of "Proper Authority and Public Declaration". Only legitimate national leaders have the competent authority to make such a declaration. At the same time, the public character reassures an occasion for national reflection concerning whether all means, except for an armed conflict, have been truly exhausted prior to the commitment of the nation to wage war This exhaustion of alternative options other than war is also the fourth standard of a just war, namely "Last Resort". The fifth element, is the "Probability of Success". Wars that do not resolve a conflict are not morally justifiable. The sixth and final element that Jackson describes for a just war is the notion of "Proportionality". The basic requirement here, is that the public good that will be obtained through the war will exceed the amount of chaos, cruelty and brutality that inevitably follows its conducting.

The above standards have formed the expectation, that every civilized nation would engage in a war only if it is assured that all of these standards are met. It stands to reason, that in the context of the "jus ad bellum", military deception could occur by a state if the latter managed to create the false impression, both to its citizens and to the other states, that all of

these standards are met, or if it creates, under deceptive actions, such a situation that will lead to the fulfillment of all of the standards above.

The second aspect of the Law of Armed Conflict is the notion of "jus in bello" or translated in English the "right (conduct) in war". Again, according to Jackson this notion "refers to a set of moral constraints on the justifiability of conduct in war" (Jackson and Fraser, 2012, p.583). Just like the "jus ad bellum" answers the question of why to wage a war, the notion of "jus in bello" answers the question of how to wage a war, and seeks to delimit violence incidental to the actual prosecution of war.

The key element of the notion of "jus in bello", is that a just war can only be fought in a just manner. There is the requirement of "proportionality" for every decision that is taken with regards to how to conduct the war. There is the need to apply the minimum force necessary, and to bring the conflict to a just and peaceful resolution as quickly as possible. These needs expand over rules governing humanitarian interests, weapons and tactics (Detter de Lupis Frankopan, 2013). All of these elements directly pertain to decisions ultimately implemented on the battlefield.

When it comes to the subcategory of tactics, military deception, which is centrally important to this thesis, is divided into two categories. These are the notions of "Perfidy" which is an unlawful deceptive action and "Ruses of War" which include deceptive acts that are considered to be lawful. In the following part of this chapter there will be an assessment regarding the existing customary and treaty provisions of the Law of Armed Conflict upon these two.

2.3 The notions of "Perfidy" and "Ruses of War" under closer consideration

The Law of Armed Conflict distinguishes military deception into two categories, "Perfidy" and "Ruses of War". In a first short explanation of these two notions, the term

"Ruses of War" refers to acts of deception that any good commander would use to confuse the enemy about the military situation, while the term "Perfidy" refers in essence to the outright treachery or the breach of good faith (The International Committee of the Red Cross, 2002). In the following paragraphs, these definitions will be analyzed in depth.

The analysis will begin with two documents of historical value, i.e. "The Lieber Code" (1863) and "The Brussels Declaration" (1874), which are the very first codifications on how to conduct war and are considered to be the ancestors of The Hague Conventions of 1899 and 1907, which will be examined in the next paragraph. "The Lieber Code" is a manual of instructions to the Union soldiers of the United States, which dictated the way that they were to behave during the battles of the Civil War. In this piece of legal history, this interesting statement can be found: Warfare "admits of deception, but disclaims acts of perfidy" (Lieber Code, 1863, art.16). In another article, the Lieber Code goes further and states that, while deception is just, necessary, and honorable, treacherous attempts to injure an enemy are reprehensible and punishable with the ultimate sanction (Lieber Code, 1863, art.101). These provisions underline that during the Civil War it was mandatory to avoid any harm through perfidious and treacherous acts, but unfortunately does not offer an exact idea of what qualifies as "perfidious" or "treacherous". Additionally, the provisions of "The Brussels Declaration" state that "Ruses of war and the employment of measures necessary for obtaining information about the enemy and the country are considered permissible" (art.14), but "the laws of war do not recognize in belligerents an unlimited power in the adoption of means of injuring the enemy" (art.12). Specifically, "murder by treachery of individuals belonging to the hostile nation or army is forbidden" (art.13). This declaration is another important contribution to the creation of today's regime and the distinction between "Ruses of War" and "Perfidy". Combatants are not allowed to use any means possible in order to injure an enemy, while it is clearly stated that murder caused by a treacherous act is

forbidden. Even though again, as in the Lieber Code, there are no tangible examples of what exactly qualifies as "Perfidy" and what as "Ruses of War".

The first international binding provisions on "Perfidy" and "Ruses of War" can be found in the "The Hague Conventions of 1899 and 1907 respecting the Laws and Customs of War on Land". More specifically, in their provisions it is stated that "Ruses of war and the employment of measures necessary for obtaining information about the enemy and the country are considered permissible" (art.24 respectively). However, "the right of belligerents to adopt means of injuring the enemy is not unlimited" (art.22 respectively) and "killing or wounding treacherously individuals belonging to the hostile nation or army is especially forbidden" (art.23 respectively). The similarity in wording and style with the Brussels Declaration, which was examined in the previous paragraph, is blatant. Unfortunately, it also suffers from the same flaws as its ancestor, since itself also does not offer a comprehensive definition that brings clarification of the acts that fall under "Perfidy" and "Ruses of War". Still, its contribution is very important, because for the first time, the international community agreed upon a common binding and codified definition.

The second binding provisions on "Ruses of War" and "Perfidy" are embodied in the Article 37 of Additional Protocol (I) to the Geneva Conventions of 12 August 1949, relating to the protection of victims of international armed conflicts" (1977). Relating to "Perfidy" the provisions of Article 37 states that:

"is prohibited to kill, injure or capture an adversary by resort to perfidy. Acts inviting the confidence of an adversary to lead him to believe that he is entitled to, or is obliged to accord, protection under the rules of international law applicable in armed conflict, with intent to betray that confidence, shall constitute perfidy" (Additional Protocol I to the Geneva Convention of 1949, 1977, art.37).

Immediately following this definition, a list of examples of "Perfidy" is given, such as "the feigning of an intent to negotiate under a flag of truce or of a surrender". The wording of the Article 37 states with clarity that except for injuring or killing an enemy, it is also forbidden to capture one by means of "Perfidy", something that was not included in Article 23 (b) of The Hague Convention analyzed previously. Moreover, Article 37 introduces three elements that need to be fulfilled in order for an act to be considered as "Perfidy": (1) The act should make the enemy believe that due to International Humanitarian Law he is either entitled protection or he is obliged not to attack, (2) the creation of trust by the enemy, either by exposing himself or by not attacking, and (3) the intentional betrayal of this trust by injuring, killing, or capturing the enemy (Watts, 2014). Regarding the notion of "Ruses of War", the wording of Article 37 offers again a detailed definition:

"Ruses of war are not prohibited. Such ruses are acts which are intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in armed conflict and which are not perfidious because they do not invite the confidence of an adversary with respect to protection under that law. The following are examples of such ruses: the use of camouflage, decoys, mock operations and misinformation" (Additional Protocol I to the Geneva Convention, 1977, art.37).

Additionally, another relevant treaty is "The Rome Statue of the International Criminal Court" (1998), which complements the Additional Protocol I (1977) to the Geneva Convention examined. Its provisions qualify as a war crime the "killing or wounding treacherously individuals belonging to the hostile nation or army" (Rome Statue, 1998, art. 8b). Of course, recognizing "Perfidy" as a war crime is of great importance, but unfortunately, the Rome Statue is influenced only by the provisions of the 1889 and 1907 Hague Conventions, since it reinstates the abstract definition of "Perfidy" by using terms

such as “treacherously”, while also recognizes the result of a perfidious act the killing or injuring of an enemy, and not the capturing of the latter.

Moving from treaty law to customary law, the List of Customary Rules of International Humanitarian Law (2005) is of great importance. The use of military deception is covered under Rules 57 to 65. More specifically, Rule 57 states that “Ruses of war are not prohibited as long as they do not infringe a rule of international humanitarian law”, while Rule 65 defines that “killing, injuring or capturing an adversary by resort to perfidy is prohibited. As in the Additional Protocol I to the Geneva Convention, the international customary law accepts actions of military deception in the format of “Ruses of War”, provided that these actions are in accordance with the Law of Armed Conflict, while prohibits not only killing and injuring, but also capturing, as a result of a perfidious act.

Through this thorough analysis of the notions of “Perfidy” and “Ruses of War” it is understood that an act of “Ruses of War” misleads the enemy but not in a way that infringes the Law of Armed Conflict and does not fall under “Perfidy”, which is an act that betrays the enemy’s trust by misusing the legal protection offered by the Law of Armed Conflict. Killing, injuring and, under the Additional Protocol I to the Geneva Convention, capturing the enemy, as a result of “Perfidy” is prohibited. The next step is to consider examples of these techniques from past wars in order to clarify even more these notions.

2.4 Military deception through the lenses of historic examples

In the first part of this sub-chapter, will be showcased examples of court cases that are relevant with the notion of “Perfidy”. To begin with, the first incident regarding “Perfidy” that will be examined is a fact that took place during the World War II. Otto Skorzeny, was the leader of the 150th SS Panzer Brigade, who created a special unit of English speaking German soldiers. This unit was supplied with uniforms and weapons of the United States

forces, which were obtained by prisoners of war, or looted from dead combatants. In this way, Skorzeny created a unit of German soldiers that looked and talked like the soldiers of the United States army. The main aim of this unit was to penetrate the lines of the United States Army and secure important checkpoints for the war's outcome, like bridges. This mission, under the code name Operation Grief, was eventually abandoned because of operation failures, but there were testimonials of United States' soldiers that German soldiers wearing United States' army uniforms had opened fire and killed their comrades. However, the Court ruling was that Skorzeny and his unit were not guilty of "Perfidy", and their actions were lawful, under the Law of Armed Conflict. This is because, there was an absence of evidence that Skorzeny and his unit, indeed engaged in combat while wearing these United States Army's uniforms. Therefore, the necessary intent to kill, or injure the enemy was lacking since they might never had entered the battle, or at least it could not be proved that they did so. The fact that they penetrated the lines of the enemy successfully, was alone not enough to lead the Court to find Skorzeny and his unit guilty of "Perfidy" (United States v. Skorzeny, 1949).

A second example regarding "Perfidy", occurred during the Gulf War in 1991. During a battle in this war, soldiers of Iraq's military forces were feigning surrendering. When soldiers of the Coalition forces approached them, other Iraqi soldiers who were hiding nearby, opened fired. In the same battle, an Iraqi officer pretended to surrender and when he came closer to the enemy drew a hidden pistol and opened fired. Coalition forces, fired back to defend themselves and shot him dead. Both incidents were considered as examples of "Perfidy", and therefore prohibited according to the Law of Armed Conflict, since there was the intent to kill the enemy (U.S. Department of Defense,1992).

Another case of "Perfidy" took place during the Bosnian War, and more specifically throughout the Srebrenica massacre in July 1995. Srebrenica was a safe haven protected by

Dutch soldiers of the United Nations. The Serbian soldiers were disguised as soldiers of the United Nations, wearing stolen U.N. uniforms and driving stolen U.N. vehicles. Driving these stolen vehicles through this free-zone, they declared that they were United Nations' peacekeepers. (Human Rights Watch, 1995). The Bosnians, believing that the Serbian soldiers were indeed United Nations' soldiers surrendered to them. After the order of Ratko Mladic, Serbian Army's General, both Bosnian fighters and Bosnian civilians that had surrendered were murdered (Rohde, 2012). In this example, the Serbian Army by pretending to be peacekeepers of the United Nations gained the Bosnians' confidence, and afterwards murdered the latter. Momir Nikolic, captain first of the Serbian Army, was sentenced for crimes against humanity to twenty-seven (27) years of imprisonment (Prosecutor v Momir Nikolic, 2002).

Regarding the notion of "Ruses of War", a clear example is Operation Fortitude during the World War II. Before D-day, the Allied Forces fooled the German army that the landing of their army would take place in Pas de Calais. Of course, the original landing was planned to take place in Normandy (Donovan, 2002). The Allied Forces managed to fake an entire army of dummies and decoys, that distracted the Germans and made them hold their positions and troops in Pas de Calais, while the real army of the Allied Forces was heading to Normand. This case is a clear example of a successful mock operation that is acceptable by the Law of Armed Conflict as an act of "Ruses of War".

2.5 Preliminary Conclusions: From the Trojan Horse to deceptive AI machines

The assessment of the use of military deception has illustrated that there are two important notions for qualifying a deceptive act as a perfidious one. Firstly, there is a need for the intent to harm the enemy (kill, injury, or capture under the Additional Protocol I), and secondly, there has to be a breach of trust. The latter trust relationship must be based on

International Humanitarian Law, and more specifically on a rule that obliges the combatants to offer protection to their enemies or to restrain from attacking them.

A conclusion from the sources that were examined above is that the provisions of "The Hague Conventions of 1899 and 1907" prohibit the treacherous killing or wounding of an enemy, while those of "The Additional Protocol I to the Geneva Convention of 1949" (1977) prohibit additionally the capturing of an enemy as a result of a perfidious act. Moreover, the "Rome Statute of the International Criminal Court" (1998) uses exactly the same wording with the "The Hague Conventions of 1899 and 1907" and qualifies as a war crime the treacherous killing or wounding of an enemy. Moving from treaty law to customary law, Rule 65 of the Customary Rules of International Humanitarian Law (2005) follows the provisions of "The Additional Protocol I to the Geneva Convention of 1949" (1977) and define as prohibited the killing, injuring, and capturing of an enemy by resort to perfidy. Thus, it is understood that killing, wounding, and capturing an enemy, as a result of a perfidious act, is prohibited and unacceptable, but only killing or injuring would qualify for a war crime.

As regards the notion of "Ruses of War", these acts are not prohibited. "The Additional Protocol I to the Geneva Convention of 1949" (1977) states in Its provisions that their intention must be to mislead the enemy or to induce him to act recklessly. Additionally, such acts shall not infringe an applicable law of the Law of Armed Conflict, and shall not be perfidious by inviting the confidence of an enemy with respect to a protection under the Law of Armed Conflict. Also, according to "The Hague Conventions of 1899 and 1907" the employment of necessary methods for obtaining information about the enemy and the country are permissible, while customary law and specifically Rule 57 of "The List of Customary Rules of International Humanitarian Law" (2005) do not prohibit such acts as long as they do not infringe a rule of international humanitarian law. Examples stated in the provisions of the

Additional Protocol I to the Geneva Convention of 1949" (1977) of "Ruses of War" are the use of camouflage, decoys, mock operations and misinformation.

Back in the myth of the Trojan Wars described in the beginning of this chapter, the Greeks breached the confidence of the Trojans with the intent to kill them. This confidence had been created because of a religious gift, which was hiding a treacherous secret. The act of the Greeks was not acceptable by the Gods, and centuries later, the Law of Armed Conflict has not changed the way that military deception is perceived as acceptable or not. Misusing the provisions of International Law to deceive the enemy with the intent of causing harm, is considered as unacceptable and is prohibited.

Chapter 3: The problems that arise from the use of deceptive AI machines on the battlefield

In the previous chapter the author delivered a thorough examination of the rules of the Law of Armed Conflict on military deception. The next step is to offer the answer to the final sub-question, i.e. "What are the problems that arise from the use of deceptive AI machines on the battlefield?". In order to do so, a deeper analysis of the rules of the Law of Armed Conflict as applied to the deceptive AI machines that were examined in the first chapter will take place. The problems that arise from the use of deceptive AI machines will be investigated, and the author will assess whether these problems can lead to challenges for the rules of the Law of Armed Conflict on military deception.

To accomplish this goal, primary sources and conclusions made in the previous chapters will be used. In addition to them, new primary and secondary sources will be examined, too. By investigating these sources, the author will examine three problematic situations, i.e. the participation of the deceptive AI machines of the state of the art in

hostilities, the element of these machines to act unexpectedly, and the responsibility issues arising if these machines proceed to a violation of a rule of the Law of Armed Conflict.

Specifically, regarding primary sources the author will analyze treaties, such as the "Additional Protocol (I) to the Geneva Conventions of 12 August 1949, relating to the protection of victims of international armed conflicts" (1977) and Its provisions on the development of new means or methods of warfare (art. 36), the combatant status (art. 43), and the State's responsibility (art.91), the "The Hague Convention" (1907) and Its provisions on the State's responsibility (art. 3), the "Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea" (1949) and Its provisions of the wounded, sick, and shipwrecked members status, and the "Protocol II of 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" (1979, as amended 1996), and Its provisions on devices which are associated with treachery and perfidy (art. 7). Moreover, customary law, i.e. Customary rules on the "combatant status" (Rule 3 & 4, of the 161 Rules of customary international humanitarian law[4]), and on "State responsibility" will be examined.

Considering secondary sources, works will be used that help in interpreting and analyzing the above mentioned primary sources, i.e. texts of Detter de Lupis Frankopan (2013), as well as reports of the Advisory Service of the International Committee of the Red Cross on International Humanitarian Law (2004), and the Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law (International Committee of the Red Cross (2009).

[4] The numbering of the Rules is according with the most authoritative and wide-ranging compilation of customary International Humanitarian Law, i.e. Jean-Marie Henckaerts' and Louise Doswald-Beck's, Customary International Humanitarian Law (Cambridge University Press, 2005), Vol. 1, Rules and Vol 2, Practice.

This thesis has a limited focus on the challenges that are created for the rules of the Law of Armed Conflict on military deception due to the use of deceptive AI machines. In view of this focus, this chapter will only consider the elements that have a direct relevance for the research. Thus, the author will exclude from his assessment other issues that are related to the use of Artificial Intelligence on the battlefield. For instance, the killing, wounding, or capturing of a combatant without the use of a deceptive tactic, or the murdering of a civilian. Moreover, possible suggestions for revising the existing Law of Armed Conflict are also outside of the scope of the central research question.

3.1 Deceptive AI machines and the conduct of hostilities

The first issue that derives from the use of deceptive AI machines in warfare is the placement of such a machine itself in the battlefield, and the question that arises is whether these machines are allowed to be in such a place.

Entering the battlefield means conducting hostilities, under the language of the Law of Armed Conflict. The acts that qualify as hostilities are any operations that are likely to affect in a negative way the military capacity/operations of a party in an armed conflict, or to cause the injury, death, destruction of persons or objects protected against direct attack (International Committee of the Red Cross, 2009). As examined in the first chapter, deceptive AI machines can perceive the world around them and based on this perception are able to perform, in an independent and unexpected way, deceptive behaviors that naturally require intelligence, such as making decisions regarding both when to deceive and how to deceive humans or other machines. Consequently, it is understood that the deceptive acts of a machine of Artificial Intelligence are able to adversely affect the armed forces of the enemy. Thus, these acts qualify as the conducting of hostilities.

However, in order to participate in the conducting of hostilities these machines need to obtain the combatant status, since only combatants have the right to direct participate in hostilities. Specifically, by virtue of the article 43, para. 2 of the Additional Protocol (I) to the Geneva Conventions (1977): "Members of the armed forces of a Party to a conflict … are combatants, that is to say, they have the right to participate directly in hostilities" (Additional Protocol I to the Geneva Convention, 1977, art. 43). Therefore, in the subsequent paragraphs the author will examine whether these machines can obtain the combatant status and participate in hostilities. If the answer to that query is a negative one, the author will examine other possible means of employing these machines in an armed force.

The starting point in responding to this question, is to examine the treaties and the customary rules that compose the Law of Armed Conflict and analyze if there is room for such an interpretation. According to paragraph 1 of Article 43 of the "Additional Protocol I to the Geneva Convention" (1977):

"the armed forces …consist of all organized armed forces, groups and units which are under a command responsible to that Party for the conduct of its subordinates. Members of the armed forces of a Party to a conflict other than medical personal and chaplain are combatants" (Additional Protocol I to the Geneva Convention, 1977, art. 43).

Additionally, under the Customary rules on the "combatant status", namely Rule 3 and Rule 4 of Customary International Humanitarian Law, "all members of the armed forces of a party to the conflict are combatants", while also "the armed forces of a party to the conflict consist of all organized armed forces, groups and units which are under a command responsible to that party for the conduct of its subordinates" (The List of Customary Rules of International Humanitarian Law", 2005, rule 3 & rule 4).

It is noticed that in these provisions the combatants are defined as "members" of the armed forces. Thus, based on the exact wording of the international treaty and customary law it appears that deceptive AI machines could qualify as combatants as long as these machines are members of an armed force and their conduct falls under the command of someone responsible for them. However, such an interpretation of the above provisions is erroneous. By taking a more careful look into the provisions of the Law of Armed Conflict, we are confronted with the understanding that the status of the "combatant" does not only include the right of direct participation in the hostilities, but instead, it also encompasses the right of the combatant to be protected when wounded, sick, lost at sea[5] or imprisoned[6]. It is understood that the status of the combatant does not only impose obligations to the one who bares it, but also offers rights and grants protection under specific situations. Thus, it is incompatible with inanimate objects, like machines.

Moreover, in contrast with Article 43 of the "Additional Protocol I to the Geneva Convention" (1977) the personal nature of combatants is clear in other primary sources of the Law of Armed Conflict, even not explicit mentioned. Specifically, according to Article 12 of the "Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea" (1949), wounded or sick members of the armed forces shall not be left without medical assistance and care, and shall not be intentionally laid open to conditions which expose them to contagions or infections (Geneva Convention II, 1949, art. 12). The personal nature of the members of the armed forces is clear in this provision and derives from the use of nouns and adjectives, such as "medical

[5] Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea" (1949, art.43).
[6] Geneva Convention (III) relative to the Treatment of Prisoners of War (1949, art.4).

assistance", "wounded" and "sick", the meaning of which is in connection with humans rather than machines.

Additionally, the conceptual core of the Law of Armed Conflict is to limit the effects of warfare and to protect the people that are not participating in the hostilities or no longer participate (Advisory Service on International Humanitarian Law, 2004). Therefore, the Law of Armed Conflict refers to people and its focal point revolves around the protection of people during warfare.

In view of the above, it us understood that the combatant status has a strong connection with humans and it is incompatible with inanimate objects, like machines. It does not only include the right of direct participation in the hostilities, but it also encompasses the right of the combatant to be protected when wounded, sick, lost at sea or imprisoned. It is understood that the combatant status is a human status, since it exists not only to impose obligations but also to protect the combatants under specific situations. Thus, deceptive AI machines cannot participate directly to the conducting of hostilities in this legal status.

However, the deceptive AI machines can be placed on the battlefield as "new means/methods of warfare" if they fall within the criteria of Article 36 of the "Additional Protocol (I) to the Geneva Conventions of 12 August 1949 (1977). Specifically, according to its provisions:

"In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party" (Additional Protocol I to the Geneva Convention of 1949, 1977, art. 36).

From the above it is clear that any Party that aims to use a new form of weapon, mean or method of warfare is obliged to ascertain that its employment will not be prohibited by any rule of international law applicable to the Party. Thus, the deceptive AI machines can enter the battlefield, if the Party that will use them determine that their employment is not prohibited by any rule of international law applicable to the Party. Specifically, for the purposes of this thesis, the deceptive AI machine must not use its deceptive tactics to kill, injury, or capture an enemy by resort to "Perfidy", since such an act is prohibited by the Article 37 of the Additional Protocol (I) to the Geneva Convention of 1949 (1977). If the deceptive AI machine meets this criterion, this means that its placement on the battlefield as a "new mean/method of conducting warfare" do not challenge the rules of the Armed Conflict on military deception.

In the following paragraphs, the author will analyze the capacities of the deceptive AI machines that were introduced to the reader under the first chapter in the light of the rules of the Law of Armed Conflict that were examined under the second chapter. The deceptive AI machines that will not be considered as perfidious will be able to enter the battlefield as a "new mean/method of conducting warfare".

3.1.1 The examined deceptive AI machines: Are their acts perfidious? As was analyzed in the first chapter, deceptive AI machines perceive the world around them and perform, in an independent and unexpected way, deceptive behaviors. These deceptive behaviors can be addressed towards humans or other machines and take various forms. In the subsequent paragraphs the author will analyze whether each of the examined deceptive AI machines has the ability to act perfidious.

In the first examined research conducted by Floreano & Keller (2010), the deceptive AI machines which were developed had the capacity to conceal information. Specifically,

these machines selected to stop emitting light when they were close to the "food area", since they were competing for the "food points", and did not want to attract other AI machines there. In this manner, they gained more "food points" for themselves. Such acts if implemented on the battlefield, could qualify as "Ruses of War", since their intention is to mislead the enemy without inviting the confidence of the latter with respect to a protection under the Law of Armed Conflict (The Additional Protocol I to the Geneva Convention of 1949, 1977, art.37.2).

In the second research, Morin et al. (2012) developed deceptive AI machines capable of implementing camouflage techniques by changing their appearance and body color in order to match their environment. Article 37 para. 2 of "The Additional Protocol I to the Geneva Convention of 1949" (1977), specifically states that examples of "Ruses of War" is the "camouflage". Consequently, such acts qualify as "Ruses of War".

In the third and the fourth researches that examined, Wagner & Arkin (2011) and Shim & Arkin (2012) managed to develop deceptive AI machines capable of misinformation. In particular, Wagner & Arkin focused on the actions, beliefs and communications of a deceptive AI machine which managed to hide from another AI machine. The hiding robot's tactic was to deliberately use false communication in order to indicate a false hiding position. In their research, Shim & Arkin developed deceptive AI machines that purposely visited locations that they had not stored their "virtual food", in order to deceive their competitors and drag them away from the locations where the food was actually stored. Article 37 para. 2 of "The Additional Protocol I to the Geneva Convention of 1949" (1977), specifically states that examples of "Ruses of War" is the "misinformation". Consequently, such acts qualify as "Ruses of War".

Lastly, in the fifth research Terada & Ito (2011) developed a deceptive AI machine which was concealing its capabilities. Specifically, during a game between the deceptive AI machine and a human competitor, the deceptive AI machine concealed from the human competitor its capability to say the special sentence in less than 5,5 seconds. Because of this deceptive behavior, the human partner was seen while moving and thus, lost the game. Such acts if implemented on the battlefield, could qualify as "Ruses of War", since the deceptive AI machine's intention is to induce the enemy to act recklessly without inviting the confidence of the latter with respect to a protection under the Law of Armed Conflict (The Additional Protocol I to the Geneva Convention of 1949, 1977, art.37.2).

Having examined the above deceptive AI machines in the light of the Laws of Armed Conflict on military deception, it is understood that the existing state of the art is such that these machines are incapable of acting in a perfidious way. Therefore, their placement on the battlefield as a "new mean/method of conducting warfare" do not challenge the rules of the Armed Conflict on military deception.

3.2 "What if?": Problems arising from the element of the deceptive AI machines to act unexpectedly

As concluded above, the deceptive AI machines of the state of the art are incapable of proceeding to perfidious acts. Their deceptive capabilities are within the limits of what is permissible under the rules of the Law of Armed Conflict on military deception. However, as was examined in the first chapter of this dissertation these machines are characterized by the element of acting unexpectedly. It was understood that a machine of Artificial Intelligence is not constrained by its creator, but instead it enjoys, at a greater or lesser degree, independent behavior that can lead to unexpected acts.

There are problems deriving from this element of the deceptive AI machines to act unexpectedly. Specifically, the question that arises is what are the tools provided by the Law of Armed Conflict and its rules on military deception if a deceptive AI machine infringe a law of the Law of Armed Conflict or proceed to a perfidious act by acting unexpectedly.

To begin with, according to Article 37 of the "The Additional Protocol I to the Geneva Convention of 1949" (1977, paragraph 2) in order a deceptive act to be considered as an act of "Ruses of War" it "shall not infringe an applicable law of the Law of Armed Conflict, and shall not be perfidious". The same wording can be found also under the customary rules of the Law of Armed Conflict on the "Ruses of War". In particular, "Ruses of War" are not prohibited "as long as they do not infringe a rule of international humanitarian law" (Rule 57 of "The List of Customary Rules of International Humanitarian Law", 2005). Therefore, it is understood that the rules of the Law of Armed Conflict clearly declare us unlawful any deceptive act that infringes an applicable law of the Law of Armed Conflict, or evolves to a perfidious act.

An additional legal tool provided by the rules of the Law of Armed Conflict on military deception that declares as unlawful any treacherous or perfidious act is the Articles 22 and 23 of "The Hague Conventions" (1899 & 1907), that were examined in the second chapter of this thesis. According to these articles, the right of belligerents to adopt means of injuring the enemy is not unlimited and especially killing or wounding treacherously individuals belonging to the hostile nation or army is forbidden. Therefore, it is clear that rules of the Law of Armed Conflict on military deception forbid the use of any mean that can lead to a treacherous wounding or killing of the enemy.

Lastly, according to article 7 of the Protocol (II) of 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" (1979, as amended 1996), it is prohibited in all circumstances to use any kind of devices which are associated with treachery and perfidy. Past examples of military equipment that was forbidden because was considered as perfidious according to the above rules of the Law of Armed Conflict, are various types of gases, such as mustard gas and phosgene gas, plastic landmines, and spike pits (Detter de Lupis Frankopan, 2013). Thus, this provision can be considered as a third legal tool to declare as unlawful a deceptive AI machine which infringes a law of the Law of Armed Conflict or proceeds to a perfidious act.

In conclusion of the above legal provisions and in connection with the first part of this chapter, the fact that a deceptive AI machine succeeds to enter the battlefield as "a new mean/method of warfare" by meeting the requirements of the Article 36 of the Additional Protocol I to the Geneva Convention of 1949 (1977), does not have the meaning that its lawfulness is not anymore questionable. Throughout its military use, a deceptive AI machine shall be lawful and not perfidious in order to continue qualify as "Ruses of War" and be acceptable.

3.3 Deceptive AI machines and the responsibility for their acts

The last problem that will be addressed regarding the use of deceptive AI machines on the battlefield is issues of "Responsibility". The question that arises is who is responsible for the acts of these machines in case they proceed to a violation of a rule of the Law Armed Conflict.

To begin with, the Rules of Customary International Humanitarian law, state that "A State is responsible for violations of international humanitarian law attributable to it, including: (a) violations committed by its organs, including its armed forces" (Rule 149 of "The List of Customary Rules of International Humanitarian Law", 2005). Deceptive AI

machines are a "new mean/method of warfare" adopted by the Party of an Armed Conflict. This Party, according to the provisions of Article 36 of the Additional Protocol I to the Geneva Convention of 1949 (1977), has the obligation to determine whether their employment would, in some or all circumstances, be prohibited by any rule of international law applicable to this Party. The armed forces of this Party are using on the battlefield these "new means/methods" of warfare. Since a State is responsible for violations of international humanitarian law committed by its armed forces, it is responsible also for the means/methods used by its armed forces. Thus, under the Rules of Customary International Humanitarian law responsible for the acts of the deceptive AI machines, in case they proceed to a violation of a rule of the Law Armed Conflict, is the State to which the armed force using these machines belongs to.

However, provisions of treaty law, and specifically the Article 3 of "The Hague Convention" (1907) and the Article 91 of the "Additional Protocol (I) to the Geneva Convention of 1949 (1977), state that a State is responsible for "all acts committed by persons forming part of its armed forces". These articles require the violation of a rule of the Law Armed Conflict to results from a person's act specifically. Therefore, these provisions cannot be used to base State's responsibility resulting from an act of a machine of Artificial Intelligence. It should be clear, that this is not only the case for a deceptive AI machine, but the case for every machine for Artificial Intelligence. Deceptive AI machines, autonomous vehicles in military operations, or autonomous weapons share the same problem of responsibility when it comes to the violation of a rule of the Law Armed Conflict. Therefore, this issue cannot be perceived as a challenge specifically for the rules of the Law of Armed Conflict on military deception, but instead it is one for the Rules of the Law of Armed Conflict on State's responsibility.

Lastly, another concern regarding responsibility is security issues. Modern warfare is technology led and the armed forces use their skills to gain unauthorized access to systems of their enemies. Thus, if a deceptive AI machine is hacked and is used for purposes that are against the Laws of Armed Conflict, issues regarding who is going to be responsible and how it will be possible to be proved that such an unauthorized access occurred, emerge. Therefore, these machines need to have security design features implemented in order to be hack-proof.

3.4 Coming closer to the conclusion

By answering the third sub-question of the problems that arise through the use of deceptive AI machines on the battlefield, the author has reached the end of his analysis. In this last chapter, three problematic situations were examined, i.e. the participation of the deceptive AI machines of the state of the art in hostilities, the element of these machines to act unexpectedly, and the responsibility issues arising if these machines proceed to a violation of a rule of the Law of Armed Conflict.

As regards the first problematic situation and in accordance with the focus of this dissertation, it is concluded that the deceptive AI machines of the state of the art can be placed on the battlefield as a "new mean/method of warfare" since they will fall within the criteria of Article 36 of the "Additional Protocol (I) to the Geneva Conventions of 12 August 1949 (1977). They are not prohibited by the rules of the Law of Armed Conflict on military deception, since all the five examples of the deceptive AI machines of the state of the art are qualified as "Ruses of War" and thus, permissible.

Considering the second problematic situation, it was understood that throughout its military use, a deceptive AI machine shall be lawful and not perfidious in order to continue to qualify as "Ruses of War". In case its element to act unexpectedly turns a lawful deceptive AI machine into a machine of unlawful and perfidious behaviors, the Law of Armed Conflict

provides via Its rules[7] the necessary tools for this machine to be declared unlawful and be removed from the battlefield.

Lastly, regarding the third problematic situation, the author considers that exists uncertainty in the Law of Armed Conflict. Specifically, based on the Customary Rules of the Law of Armed Conflict (Rule 149 of "The List of Customary Rules of International Humanitarian Law", 2005) the State in which the armed force using the deceptive AI machines violating a rule of the Law Armed Conflict belongs to, is responsible for this violation. But, according to the respective provisions of the treaty law of the Law of Armed Conflict (Article 3 of "The Hague Convention", 1907 and Article 91 of the "Additional Protocol (I) to the Geneva Convention of 1949", 1977) the State cannot be responsible because provided by law the violation must result from a person's act. However, this uncertainty cannot be perceived as a challenge specifically for the rules of the Law of Armed Conflict on military deception, but instead it is one for the Rules of the Law of Armed Conflict on State's responsibility, since the same issue arise for any machine of Artificial Intelligence, such as Deceptive AI machines, autonomous vehicles in military operations, or autonomous weapons, that violate a rule of the Law of Armed Conflict. Last but not least, the probability of deceptive AI machines to be hacked should be taken also into account.

. Considering the above the author is ready to give the answer to the main research question of this thesis.

[7] namely, Article 37.2 of the "The Additional Protocol I to the Geneva Convention of 1949", 1977, the Articles 22 and 23 of "The Hague Conventions",1899 & 1907, and Article 7 of the "Protocol (II) of 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", 1979 as amended 1996.

Final Chapter: Conclusion and Recommendations

Conclusion

The author reaches the end of this analysis by recalling the main goal of the research, as was formulated in the very beginning. The goal of this thesis was to examine whether there are possible frictions between the use of deceiving machines of Artificial Intelligence and the Law of Armed Conflict. This was done by firstly analyzing the capacities of the deceptive AI machines, secondly the rules of the Law of Armed Conflict on military deception, and thirdly the problematic situations that arise. By setting the limits of this thesis, was recognized that only elements that have a direct relevance for this research will be examined, excluding from the assessment other situations that can emerge from the use of Artificial Intelligence on the battlefield, such as the killing, wounding, or capturing of a combatant without the use of a deceptive tactic, or murdering of a civilian. Moreover, possible suggestions for revising the existing Law of Armed Conflict were also outside of this dissertation's scope. The goal of this thesis was specifically expressed into the research question:

"Does the use of deceptive AI machines in warfare challenge the rules of the Law of Armed Conflict on military deception, and if it does what are these challenges?"

In answering this question, the author divided it into three sub-questions: 1. "What is a deceptive AI machine?", 2. "What is permissible and what is forbidden regarding the use of deception in warfare under the rules of the Law of Armed Conflict", and 3. "What are the problems that arise from the use of deceptive AI machines on the battlefield?". Each of these sub-questions was the focal point for a chapter of this thesis.

Thus, in the first chapter it was understood that deceptive AI machines perceive the world around them and based on this perception are able to perform, in an independent and unexpected way, deceptive behaviors. In the second chapter, by analyzing in depth primary

and secondary sources of the Law of Armed Conflict, was realized that military deception is distinguished into two categories, i.e. “Perfidy”, which refers in essence to outright treachery or the breach of good faith and is considered to be a war-crime, and “Ruses of War”, which refers to acts of deception that any commander can use to confuse the enemy regarding the military situation. Lastly, in the third chapter, the deceptive AI machines of the state of the art were analyzed in the light of the Law of Armed Conflict, and three problematic situations were examined, i.e. the participation of the deceptive AI machines of the state of the art in hostilities, the element of these machines to act unexpectedly, and the responsibility issues arising if these machines proceed to a violation of a rule of the Law of Armed Conflict. Having answered the sub-questions of each chapter, and accessed the respective conclusions, the author will hereby deliver the answer to the main question of the thesis:

The use of deceptive AI machines in warfare does not challenge the existing rules of the Law of Armed conflict on military deception.

Specifically, the answer is negative because, the rules of the Law of Armed Conflict on military deception enjoy a high degree of flexibility. By setting clearly the criteria according to which deceptive tactics can be considered as permissible or as prohibited It has managed to adapt from the early years of the 20th Century until today, 2017, to a wide variety of developments regarding deceptive means and tactics employed in the battlefield. The dynamism of Its corpus can be seen not only in the provisions of treaties such as “The Hague Conventions of 1899 and 1907”, The Additional Protocol I to the Geneva Convention of 1949” (1977), and the “Rome Statute of the International Criminal Court” (1998), but also to the wording of the Customary International Humanitarian Law. The elements for qualifying a deceptive tactic as perfidious and prohibited are clearly defined. Firstly, there is a need for the intent to harm the enemy (kill, injury, or capture under the Additional Protocol I), and secondly, there has to be a breach of trust. The latter trust relationship must be based on a

provision of International Humanitarian Law, that obliges to offer protection or to restrain from attack. Any new deceptive weapons, tactics, methods or means that meet these criteria are forbidden, while on the contrary any of them that are intended to mislead the enemy or to induce him to act recklessly without inviting the confidence of the latter with respect to a protection under the Law of Armed Conflict, and without infringing a rule of international humanitarian law, are acceptable and qualify as "Ruses of War". So, deceptive AI machines that qualify as perfidious are forbidden, and those of them that do not qualify as perfidious are acceptable and permissible. Thus, deceptive AI machines do not challenge the rules of the Law of Armed Conflict on military deception.

Moreover, the answer is negative, because the Law of Armed Conflict has proved to be resilient through the principles that define Its operations. The key element of "jus in bello", is that a just war can only be fought in a just manner. There is the requirement of "proportionality" for every decision that is taken with regards to how to conduct the war. There is the need to apply the minimum force necessary, and to bring the conflict to a just and peaceful resolution as quickly as possible. These needs expand over rules governing interests, weapons and tactics and serve as a tool for defying standards of what types of technological developments are lawful and accepted for conducting warfare. If deceptive AI machines do not meet these standards, they are considered to be unlawful and unacceptable, and therefore they do not challenge the rules of the Law of Armed Conflict on military deception.

Lastly, the answer is negative, because the Law of Armed Conflict sets clearly through Article 36 of the "Additional Protocol (I) to the Geneva Conventions of 12 August 1949" (1977) the requirements that a new weapon, mean, or method needs to fulfil in order to be used by an armed force on the battlefield. Specifically, the Party in which this armed force using the new weapon, mean, or method, belongs to, is obliged to ascertain that their

employment will not be prohibited by any rule of international law applicable to the Party. So, if the Party cannot prove that the deceptive AI machine in question is not prohibited by any rule of international law applicable to that Party, the deceptive AI machine is unlawful and cannot be used by the armed forces of the Party. Thus, the deceptive AI machines do not challenge the rules of the Law of Armed Conflict on military deception.

As regards the element of AI machines to act unexpectedly and the probability a lawful deceptive AI machine to turn into an AI machine of unlawful and perfidious behaviors the Law of Armed Conflict provides a high level of protection. Through the Article 37.2 of the "The Additional Protocol I to the Geneva Convention of 1949" (1977), the Articles 22 and 23 of "The Hague Conventions" (1899 & 1907), and the Article 7 of the "Protocol (II) of 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" (1979 as amended 1996) the Law of Armed Conflict provides the tools for such a AI machine to be declared unlawful and be removed from the battlefield.

The uncertainty that was expressed by the author regarding State's responsibility for violations of the Law of Armed Conflict by deceptive AI machines, cannot be perceived as a challenge specifically for the rules of the Law of Armed Conflict on military deception. Instead it is one for the Rules of the Law of Armed Conflict on State's responsibility, since the same issue arise for any machine of Artificial Intelligence, such as Deceptive AI machines, autonomous vehicles in military operations, or autonomous weapons, that could violate a rule of the Law of Armed Conflict. However, this challenge is at the same time a proof of the resilience of the Law of Armed Conflict in general, since thanks to this uncertainty AI machines that pose high risks of violating rules of the Law of Armed Conflict, such as autonomous weapons, are not allowed to be deployed by armed forces.

In this point and before the end of the conclusion, the author will acknowledge the limitations of the research. The use of the methodology of Doctrinal (Black Letter) legal analysis was an important step to recognize in depth the primary and secondary sources of the Law of Armed Conflict and assess the legality of the deceptive AI machines. However, the luck of a critical evaluation based also on non-legal sources limits the production of a more creative work. The author did not proceed to a combined methodology because he would need more amount of time in order to adequately understand and process information from other fields of studies, and this was not an option based on the specific time limitations for a master thesis. Nevertheless, the doctrinal focus was a good starting point, and in future researches on a wider topic the author can use a more creative methodology in order to provide alternative and different perspectives.

Recommendations

My personal opinion as a law student in the field of Law and Technology is that complex notions like the deceptive AI machines, create complex issues that require synergy and collaboration of scholars from different fields of studies in order to be addressed. Therefore, this paragraph is an open call to computer scientists who are conducting research in the field of robotic deception to join our forces, knowledge, and passion in order to deliver a multidisciplinary and thorough analysis on similar issues.

References

Ackerman, E. (2012). Georgia Tech Robots Learn Deceptive Behaviors from Squirrels. IEEE Spectrum: Technology, Engineering, and Science News. Retrieved 15 February 2017, from http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/robots-learn-deceptive-behaviors-from-squirrels

American Society of International Law (2013), Electronic Resource Guide written by Policastri J. & Stone S.D. Available at https://www.asil.org/sites/default/files/ERG_International%20Humanitarian%20Law%20%28test%29.pdf,

Arkin, R., & Shim, J. (2013). A Taxonomy of Robot Deception and its Benefits in HRI. In 2013 IEEE International Conference on Systems, Man, and Cybernetics.

Arkin, Ronald C. (2011) The Ethics of Robotic Deception, Mobile Robot Laboratory, Georgia Institute of Technology, page 1, http://www.cc.gatech.edu/ai/robot-lab/online-publications/deception-final.pdf

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

Beyer, R., & Sayles, E. (2015). The Ghost Army of World War II: How One Top-Secret Unit Deceived the Enemy with Inflatable Tanks, Sound Effects, and Other Audacious Fakery (1st ed.).

Boden, M. (1987). Artificial intelligence and natural man (first edition). London: The MIT Press.

Bond, C., & Robinson, M. (1988). The evolution of deception. Journal Of Nonverbal Behavior, 12(4), 295-307. http://dx.doi.org/10.1007/bf00987597

Brussels Declaration (1874), Project of an International Declaration concerning the Laws and Customs of War. Brussels, 27 August 1874

Canine member of the Armed Forces Act (2012), H.R. 4103 and S. 2134, https://awionline.org/content/canine-members-armed-forces-act accessed 17.04.2017

Carson, Thomas L. (2009) Lying, Deception and Related Concept, The Philosophy of Deception. ed Clancy Martin, New York: Oxford University Press, pp 153-187.

Clark M.H., & Atkinson D. (2013) (Is there) A Future for Lying Machines? Deception & Counter-Deception Symposium, International Association for Computing and Philosophy 2013 Conference, At College Park.

Convention (II) (1949) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea. Geneva, 12 August 1949

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 (1979, as amended 1996), Protocol on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices as amended on 3 May 1996

Crootof, R. (2014). The Killer Robots Are Here: Legal and Policy Implications. 36 Cardozo L. Rev. 1837. Available at SSRN: https://ssrn.com/abstract=2534567

Crootof, R. (2015). War, Responsibility, and Killer Robots. North Carolina Journal of International Law and Commercial Regulation, Vol. 40, No. 4, 2015. Available at SSRN: https://ssrn.com/abstract=2569298

Cummings, M. (2017). Artificial Intelligence and the Future of Warfare, page 6, International Security Department, US and the Americas Programme. www.chathamhouse.org/publication/artificial-intelligence-and-future-warfare#sthash.FscFUJ4s.dpuf

D. J. Sexton (1986) "The theory and psychology of military deception," Suny press.

Detter de Lupis Frankopan, I. (2013). The Law of War (1st ed., p. 160). Ashgate Publishing Group.

Doncieux, S., Bredeche, N., Mouret, J., & Eiben, A. (2015). Evolutionary Robotics: What, Why, and Where to. Frontiers In Robotics And AI, 2. http://dx.doi.org/10.3389/frobt.2015.00004

Donovan, M. (2002). Strategic deception (1st ed.). Carlisle Barracks, PA: U.S. Army War College.

Fahlman S. (2015). Position Paper: Knowledge-Based Mechanisms for Deception. Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems, 2015

Floreano D, Keller L (2010) Evolution of Adaptive Behaviour in Robots by Means of Darwinian Selection. PLoS Biol 8(1): e1000292. https://doi.org/10.1371/journal.pbio.1000292

Frankel, R. (2014). Dogs at War: Judy, Canine Prisoner of War. News.nationalgeographic.com. Retrieved 29 May 2017, from http://news.nationalgeographic.com/news/2014/05/140518-dogs-war-canines-soldiers-troops-military-japanese-prisoner/

Handel, Michael I. (2006). Masters of War: Classical Strategic Thought (3rd rev. and expanded ed.). London: Routledge. ISBN 0714650919.

Homer (n.d.). The Iliad. (1st ed., pp. , Epit..5.16). Apollodorus, Library and Epitome (English translation) http://perseus.uchicago.edu/perseuscgi/citequery3.pl?dbname=GreekTexts&query=Apollod.%20Epit.5.15&getid=1

Homer (n.d.). The Odyssey, Homer.

House of Commons, Science and Technology Committee (2016). Robotics and artificial intelligence, Fifth report of Session 2016-17, Transpolitica, page 6,. Retrieved from https://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf

Human Rights Watch (1995). The Fall of Srebrenica and the Failure of U.N. Peacekeeping. D713, available at: http://www.refworld.org/docid/3ae6a7d48.html

Hutchins, A., Cummings, M., Draper, M., & Hughes, T. (2015). Representing Autonomous Systems' Self-Confidence through Competency Boundaries. Proceedings Of The Human Factors And Ergonomics Society Annual Meeting, 59(1), 279-283. http://dx.doi.org/10.1177/1541931215591057

International Committee of the Red Cross (2004). Advisory Service on International Humanitarian Law. What is International Humanitarian Law?. Retrieved 17 April 2017, from http://www.icrc.org/eng/assets/files/other/what_is_ihl.pdf

International Committee of the Red Cross (2005), Customary International Humanitarian Law, Volume I: Rules, available at: http://www.refworld.org/docid/5305e3de4.html

International Committee of the Red Cross (2009). Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law. International Review Of The Red Cross, p.16. http://dx.doi.org/10.1017/s1816383109000319

International Criminal Tribunal for the former Yugoslavia, Appeals Chamber Kordić and Čerkez Case. (2004). Customary IHL - Practice Relating to Rule 3. Definition of Combatants. https://ihl-databases.icrc.org/customary-ihl/eng/docs/v2_rul_rule3

Jackson, D., & Fraser, K. (2012). Encyclopedia of Global Justice, Edited by Deen K. Chatterjee. Reference Reviews, (7), 581-584 http://dx.doi.org/10.1108/09504121211270889

Jenks, C (2010). Law from Above: Unmanned Aerial Systems, Use of Force, and the Law of Armed Conflict (March 12, 2010). North Dakota Law Review, Vol. 85, p. 649, 2010. Available at SSRN: https://ssrn.com/abstract=1569904

Kastan, B. N. (2013). Autonomous Weapons Systems: A Coming Legal 'Singularity'? (April 10, 2012). 2013 Journal of Law, Technology and Policy 45. Available at SSRN: https://ssrn.com/abstract=2037808

L. Hawthorne, (2006) "Military deception," Joint Publication, JP 3-13.4.

Lieber Code (1863), Article 16, Prepared by Francis Lieber, promulgated as General Orders No. 100 by President Lincoln, 24 April 1863, http://avalon.law.yale.edu/19th_century/lieber.asp#sec1 accessed 14.03.2017

Lieblich, E. & Benvenisti, E. (2014). The Obligation to Exercise Discretion in Warfare: Why Autonomous Weapon Systems are Unlawful (August 13, 2014). Autonomous Weapons Systems: Law, Ethics, Policy 244 (Nehal Bhuta et al, eds., Cambridge University Press, 2016, Forthcoming). Available at SSRN: https://ssrn.com/abstract=2479808

Lynch, Michael P. (2009) Deception and the Nature of Truth, The Philosophy of Deception. ed. Clancy Martin, New York: Oxford University Press, pp 188-200.

Marra, W. & McNeil, S. (2013). Understanding 'The Loop': Regulating the Next Generation of War Machines (May 1, 2012). Harvard Journal of Law and Public Policy, Vol.

36, No. 3, 2013. Available at SSRN: https://ssrn.com/abstract=2043131 or http://dx.doi.org/10.2139/ssrn.2043131

Matthew J. Greer (2015), Redefying Perfidy, Georgetown Journal of International Law, p.262

Mattox, J. M. (1998). The Ethics of Military Deception. Master Thesis At Faculty Of U.S. Army Command And General Staff College, 33-35.

McCarthy, J. (1959). In Mechanisation of Thought Processes, Proceedings of the Symposium of the National Physics Laboratory, pages 77-84, London, U.K., Her Majesty's Stationery Office.

Naval Science & Technology Strategy- Office of Naval Research. Onr.navy.mil. Retrieved 15 February 2017, from https://www.onr.navy.mil/en/About-ONR/science-technology-strategic-plan

Nilsson, N.J. (2010). The Quest for Artificial Intelligence: A History of Ideas and Achievements, page 13, Cambridge University Press.

Northrop Grumman (2006). Global Hawk capabilities Retrieved 17 April 2017 from http://www.northropgrumman.com/Capabilities/GlobalHawk/Pages/default.aspx

Patel, A. (2017). AI techniques. Theory.stanford.edu. Retrieved 15 February 2017, from http://theory.stanford.edu/~amitp/GameProgramming/AITechniques.html

Poole, D., Mackworth, A. K., and Goebel, R. (1998). Computational intelligence: A logical approach. Oxford University Press.

Prosecutor v Momir Nikolic (2002), IT-02-60/1-T (Dec. 2, 2003) para 183. http://www.icty.org/case/dragan_nikolic/4#tdec

Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I) (1977)

Rajaraman, V. (2014). JohnMcCarthy — Father of artificial intelligence. Resonance, 19(3), 198-207. http://dx.doi.org/10.1007/s12045-014-0027-9

Rich, E. and Knight, K. (1991). Artificial Intelligence (second edition), McGraw-Hill

Rohde, D. (2012). Endgame, The Betrayal and Fall of Srebrenica, Europe's Worst Massacre Since World War II (1st ed.). New York: Penguin Books.

Rome Statue of the International Criminal Court (1998), Article 8b, xi

Russell, S., & Norvig, P. (2016). Artificial intelligence: A Modern Approach, Preface page 8, (third edition), Prentice Hall Series in Artificial Intelligence

S. A. Morin, R. F. Shepherd, S. W. Kwok, A. A. Stokes, A. Nemiroski, and G. M. Whitesides, (2012) “Camouflage and Display for Soft Machines,” Science, vol. 337, no. 6096, pp. 828–832.

Schmitt, M. (2011). Drone attacks under the jus ad bellum and jus in bello (1st ed.)

Sean Watts (2014). Law-of-War Perfidy 219 MILITARY LAW REVIEW 106, 107–8, https://www.loc.gov/rr/frd/Military_Law/Military_Law_Review/pdf-files/219-spring-2014.pdf

Shim J., Arkin R.C. (2012) Biologically-Inspired Deceptive Behavior for a Robot. In: Ziemke T., Balkenius C., Hallam J. (eds) From Animals to Animats 12. SAB 2012. Lecture Notes in Computer Science, vol 7426. Springer, Berlin, Heidelberg https://link.springer.com/chapter/10.1007/978-3-642-33093-3_40

Solis, G. (2011). The Law of Armed Conflict: International Humanitarian Law in War. Journal Of Conflict And Security Law, 469.

Terada, K., & Ito, A. (2011). Can Humans Be Deceived by a Robot?-Unexpected Behavior is a Cue for Attributing Intentions to a Robot-. Journal Of The Robotics Society Of Japan, 29(5), 445-454. http://dx.doi.org/10.7210/jrsj.29.445

The Dickin Medal. PDSA. Retrieved 12 April 2017, from https://www.pdsa.org.uk/what-we-do/animal-honours/the-dickin-medal

The Hague Convention (II) respecting the Laws and Customs of War on Land and its annex: Regulations concerning the Laws and Customs of War on Land. (1899) The Hague, 29 July 1889, Article 23(b)

The Hague Convention (IV) respecting the Laws and Customs of War on Land and its annex: Regulations concerning the Laws and Customs of War on Land. (1907) The Hague, 18 October 1907, Article 23(b)

The International Committee of the Red Cross (2002), The Law of the armed conflict, Conduct of Operations, Part B, p. 4

The List of Customary Rules of International Humanitarian Law. (2005). Jean-Marie Henckaerts' and Louise Doswald-Beck's, Customary International Humanitarian Law (Cambridge University Press, 2005), Vol. 1, Rules and Vol 2, Practice.

U.S. Dep't Of Defense, Conduct Of The Persian Gulf War: Final Report to Congress O-21. (1992). http://www.dtic.mil/dtic/tr/fulltext/u2/a249270.pdf

U.S. Department of Defense (2009) FY2009-2034 Unmanned Systems (Washington, DC: US Government Printing Office) Retrieved 15 April 2017, from http://www.acq.osd.mil/sts/docs/Unmanned%20Systems%20Integrated%20Roadmap%20FY2011-2036.pdf

United States Department of Defense (2009). United States Army (2009), Unmanned Aircraft Systems (UAS) Operator (15W) http://www.goarmy.com/careers-and-jobs/browse-career-and-job-categories/transportation-and-aviation/unmanned-aerial-vehicle-operator.html

United States v. Skorzeny (1949). General Military Government Court of the US Zone of Germany

W. J. Meehan (1998) "Fm 90-2 battlefield deception," Army Field Manuals.

Wagner, A.R. & Arkin (2011) R.C. Int J of Soc Robotics https://link.springer.com/article/10.1007/s12369-010-0073-8

Winston, P. H. (1992). Artificial Intelligence (third edition). Addison-Wesley.

Zittrain, J. (2017). Openness and Oversight of Artificial Intelligence. Retrieved 15 February 2017, from https://cyber.harvard.edu/node/99783