Autonomous Weapons Systems: The Permissible Use of Lethal Force, International Humanitarian Law and Arms Control

by

Carmen Kendell Herbert

Thesis presented in fulfilment of the requirements for the degree of Master of Arts in the Faculty of Philosophy at Stellenbosch University

Supervisor: Dr Tanya de Villiers-Botha

December 2017

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

December 2017

Abstract

This thesis examines both the ethical and legal issues associated with the use of fully autonomous weapons systems. Firstly, it addresses the question of whether or not an autonomous weapon may lawfully use lethal force against a target in armed conflict, given the constraints of International Humanitarian Law, and secondly, the question of the appropriate loci of responsibility for the actions of such machines. This dissertation first clarifies the terminology associated with autonomous weapons systems, which includes a discussion on artificial intelligence, the difference between automation and autonomy, and the difference between partially and fully autonomous systems. The structure is such that the legal question of the permissible use of lethal force is addressed first, which includes discussion on the current International Humanitarian Law requirements of proportionality and distinction. Thereafter a discussion on potential candidates for responsibility (and consequentially liability) for the actions of autonomous weapons that violate the principles of International Humanitarian Law follows. Addressing the aforementioned questions is critical if we are to decide whether to use these weapons and how we could use them in a manner that is both legal and ethical. The position here is that the use of autonomous weapons systems is inevitable, thus the best strategy to ensure compliance with International Humanitarian Law is to forge arms control measures that address the associated issues explored in this dissertation. The ultimate aim in asking the associated legal and ethical questions is to bring attention to areas where the law is currently underequipped to deal with this new technology, and thus to make recommendations for future legal reform to control the use of autonomous weapons systems and ensure compliance with the existing principles of International Humanitarian Law.

Opsomming

Hierdie tesis ondersoek die etiese sowel as die regskwessies wat met die gebruik van ten volle outonome wapenstelsels verband hou. In die besonder handel dit oor die vraag of ’n outonome wapen regmatig dodelike geweld teen ’n teiken in gewapende konflik mag gebruik in die lig van die beperkinge van die internasionale humanitêre reg, sowel as die vraag oor by wie die verantwoordelikheid vir die aksies van sulke masjiene behoort te berus. Hierdie verhandeling begin deur die terminologie op die gebied van outonome wapenstelsels te verklaar, wat insluit ’n bespreking van kunsmatige intelligensie, die verskil tussen outomatisasie en outonomie, en die verskil tussen gedeeltelik en ten volle outonome stelsels. Wat struktuur betref, kom die regsvraag oor die toelaatbare gebruik van dodelike geweld eerste aan bod, met inbegrip van ’n bespreking van die huidige vereistes van proporsionaliteit en onderskeid ingevolge die internasionale humanitêre reg. Daarna volg ’n bespreking van moontlike kandidate vir verantwoordelikheid (en gevolglik aanspreeklikheid) vir die aksies van outonome wapens wat internasionale humanitêre regsbeginsels skend. ’n Ondersoek na hierdie vraagstukke is noodsaaklik om te besluit of ons hierdie wapens enigsins behoort te gebruik, en of ons dit op ’n regmatige sowel as ’n etiese manier kan gebruik. Die standpunt in hierdie verband is dat die gebruik van outonome wapenstelsels onafwendbaar is, en dus is die beste strategie om wapenbeheermaatreëls in te stel om die verbandhoudende kwessies wat in hierdie verhandeling verken word, die hoof te bied. Die einddoel met die verkenning van die verbandhoudende regs- en etiese vraagstukke is om die aandag te vestig op gebiede waar die reg tans onvoldoende toegerus is om hierdie nuwe tegnologie te hanteer, en om dus aanbevelings te doen vir toekomstige regshervorming om die gebruik van outonome wapenstelsels te beheer en voldoening aan bestaande internasionale humanitêre regsbeginsels te verseker.

Table of Contents

1. Introduction

The development of artificial intelligence [1] has long been the goal of modern technological science. Historically, most technological innovation has been driven by the military, so it is likely that one of the first forms of artificial intelligence systems will be weaponized. Weapons systems are already becoming increasingly automated, and human involvement in the lethal decision-making loop is becoming more diminished. Consider the increasing automation of warfare over history; in the 20 or so years from World War Two to the Vietnam War the amount of manpower the United States Air Force required to hit a target was reduced by 99.4%, as a result of increasing weapon targeting accuracy (Harris, 2012: 1). By the time the United States went to war with Afghanistan and Iraq in 2001 and 2003 respectively, the pilots were increasingly not even inside the plane (*ibid.* 2). As aircraft weapons have become more precise, humans have become less essential to the conduct of war, and in all probability there will eventually be a single mission commander who will control, or perhaps passively observe, a swarm of autonomous Unmanned Aerial or Ground Vehicles (UAVs and UGVs respectively) (*ibid*. 2). Weapons that can identify, track, and engage targets without human input already exist and are in use, although currently only in a defensive capacity (US Department of Defense, 2012). The development of fully autonomous weapons systems and the increasing use of UAVs and UGVs in military and para-military settings has sparked a debate about the ethical and legal use of this technology (Cummings, 2014: 1). This thesis will examine some of these questions and in order to develop an argument for arms control regulation.

There are strong but diverging views on the use of autonomous weapons systems. Proponents argue that the utilization of autonomous weapons systems would reduce the danger to soldiers' lives and cut military spending. They also maintain that autonomous weapons are able to process data far more rapidly than a human soldier, and would not be inclined to act out of fear or anger as humans would, all of which are desirable qualities within the context of war. Opponents argue that these systems lack compassion and empathy, which are important

[1] Artificial Intelligence theory will be discussed in depth in the next chapter.

inhibitors to killing needlessly, and would not possess the human judgement necessary to make the subjective assessments required by International Humanitarian Law. Opponents also express concern regarding the unclear nature of responsibility for the acts committed by autonomous agent (Human Rights Watch, 2014: 6). Due to the controversial nature of this technology, some dismiss the marriage between artificial intelligence and weapons systems outright, claiming that it is immoral for a machine to have lethal decision making power. This is a sentiment embodied by the Future of Life Institute, for example, which seeks to ban weaponized artificial intelligence (Future of Life Institute, 2015). It is my position that an outright ban of such weapons is not desirable or even possible. My reasons for this position are as follows:

Firstly, we should not seek to ban this technology since it could offer huge benefits[2] to humanity, thus it is our duty to consider its development and implementation thoroughly. Secondly, even if we were to opt for a ban, it would be difficult to enforce globally, which becomes obvious if one examines the history of the ban on antipersonnel landmines, nuclear proliferation, and even digital copyright management. The aforementioned "banned" weapons are still in existence and are in some instances (in the case of antipersonnel landmines, for example) still in production. There have been numerous efforts[3] to curb the proliferation of cruise missiles, for example, since their inception thirty odd years ago, and the limited success thereof demonstrates the difficulty of reaching arms control consensus even on weapons that have low levels of autonomy (Gormley, 2009). Authors like Peter Asaro and Robert Sparrow, despite having a negative perception of autonomous weapons systems, all propose measures for arms control. Thus, even the strongest critics recognise the inevitability of the adoption of autonomous weapons in armed conflict.[4] Therefore, the question is not whether we should

[2] Such benefits are explicitly discussed in Chapter 3, but all amount to increasing safety and security for soldiers and a reduction in collateral damage.

[3] The struggle to forge a lasting arms control measure over cruise missiles is illustrated by the listing the various, partially successful treaties aimed at dealing with the issue: 1987 The Missile Technology Control Regime,1987 Intermediate-Range Nuclear Forces Treaty, 1998 Commission to Assess the Ballistic Missile Threat, 2002 International Code of Conduct against Ballistic Missile Proliferation (Gormley, 2009). These treaties have only been partially successful because of uneven execution of controls by various states due to weak international norms (*ibid*).

[4] Fighting amounts to armed conflict if organized armed groups fight each other with a certain amount of intensity (O'Connell, 214: 229).

adopt autonomous weapons in armed conflict, but rather which restrictions we should impose on their use and what existing legal mechanisms can be used as a guideline to do this.

Formalising arms control measures is the most reasonable way forward. However, this task is complicated by the various levels of autonomy in weapons systems; we already have a variety of semi-autonomous and passive-autonomous weapons in use (Cummings, 2014: 2).[5] What is also apparent is that there is a lack of consensus over how to define and categorise autonomy regarding weapons systems. Definitions vary among the International Human Rights Council, the United Nations Security Council, and even within various branches of the United States military.[6] Finding a concrete definition is important because without such unanimity regulation, and thus any form of arms control, becomes difficult, as the organisations deploying these weapons could argue that their weapons are not bound by the international community's understanding of what autonomous weapons entail. Additionally, there are no clear criteria for what constitutes a proportional application of force.[7] Without a clearer understanding of what it entails it is difficult to establish when and if the use of force is permissible and programming autonomous weapons systems (AWS) with the concept of proportionate response becomes erratic. Another matter that needs to be addressed is the lack of recognition in International Law between differences in culpability of mental states, specifically between intentional war criminals, and those who are merely negligent or reckless. This distinction is important, since it is likely that most humans who deploy autonomous weapons that contravene International Law will be negligent rather than malicious; the difference between these two mental states determines the degree to which they are culpable. These matters need to be addressed, not only for the sake of the autonomous weapons debate but also to ensure consistent applications of proportionality by human soldiers, and to account for differences in mental culpability in international law.

[5] The differences between the levels of autonomy will be made clear in the chapter defining autonomous weapons systems.
[6] These are the most relevant voices in any discussion on armed conflict in international community. Their relevance is discussed later in Chapter 3.
[7] Proportionality is a requisite for permissible lethal force, and is fully discussed in chapter four.

Some legal scholars have argued that militaries and countries would not want to use weapons that are difficult to define, predict, and hold accountable, and that an excessive investigation into the ethical use of autonomous weapons is therefore not necessary, but this claim is speculative at best, and unlikely to materialise after the benefits of using autonomous weapons become more apparent (Asaro, 2016: 12). Asaro also believes that once states and militaries see the advantages of these kinds of weapons, they may be reluctant to regulate them (*ibid*). This is one of the reasons why it is important to define and regulate autonomous weapons before there is motivation to delay such legislation.

It is also important to adequately regulate autonomous weapons systems for a further reason. Humans have been engaged in conflict throughout recorded history, and every time we develop new weapons technology, the standards regulating that technology are set after the weapon has already been employed in armed conflict (O'Connell, 2014: 224). Only after the use of mustard gas grenades, landmines, atomic bombs, drones and so forth, have we been prompted to raise moral questions about the nature of these weapons. The efforts to ban antipersonnel landmines are probably the best example of why international consensus is important; despite the 1997 Mine Ban Treaty being one of the most widely accepted treaties in existence, large and powerful militarized countries like the United States, China and Russia, amongst other United Nations member states, are not party to the Treaty, and in fact continue to produce land-mines (International Campaign to Ban Landmines, 2016). Yet, there is a general consensus that the use of victim-triggered weapons does not meet the requirements of proportionality and distinction set out in International Humanitarian Law.[8] Without this consensus and the treaty, many more countries would still be using and producing landmines. This consensus is also putting increasing pressure on the states that use and produce landmines to stop doing so. Regulating autonomous weapons will similarly curb their proliferation. With autonomous weapons systems, given that they can think for themselves and can potentially re-programme their own protocols, it is important to set standards before the fact rather than in hindsight, lest we create weapons we cannot control. Finally, without legal intervention,

[8] International Humanitarian Law is the branch of international law that governs conduct in armed conflict, and thus dictates which kinds of weapons are permissible to use. It is more thoroughly discussed in chapter 4.

scientists may continue to develop weapons that take the kill decision further away from the humans who bear responsibility for its use (O'Connell, 2014: 228).

The first topic in this dissertation is an introduction to artificial intelligence. Understanding the basic theories surrounding artificial intelligence and the debate about whether or not it is even possible to engineer intelligence, makes understanding the arguments for and against autonomous weapons easier. The two principal artificial intelligence theorists I will examine are Alan Turing and John Searle. I will also establish the case for arguing that the internal mental states, or intentionality, of the artificial agent does not matter for establishing the permissibility of the use of lethal force and is only relevant in some (but not all) cases of responsibility. After clarifying what is meant by artificial intelligence, I will discuss what constitutes an autonomous weapon system, based on the understanding and directive on the matter put forward by the United States Department of Defense, and define related terminology. I will also discuss some of the potential benefits highlighted by proponents of autonomous weapons systems, which are essentially the goals developers of autonomous weapons bear in mind when engineering them. Lastly, I will examine some criticisms levelled by opponents against autonomous weapons systems, which will serve to illuminate what challenges, legal and ethical, need to be overcome before such weapons are deployed.

Thereafter, two chapters will be devoted to scrutinizing the two biggest obstacles proponents of autonomous weapons have to overcome. Firstly, a chapter will be dedicated to examining when the use of lethal force is permissible, specifically with regard to autonomous weapons. Currently, the application of lethal force in international armed conflict is regulated by the laws of war, which are set out in certain subsections of International Humanitarian Law. The focus will be on international law, since norms established here herald the norms that are established in national laws. One would assume the requirements for permissible killing are the same for autonomous weapons as for humans; however, part of the reason war-time killing is deemed permissible is because it is a human agent that triggers the attack (O'Connell, 2014: 225). Thus the current legal framework needs to be reviewed and possibly revised in order to cater for the development of this new technology. The question is whether it is possible for an autonomous weapon to adhere to the legal framework established in the laws of war as well as

a human agent. The sentiment of International Humanitarian Law is overtly humanitarian, so any autonomous weapons system must be compliant with the humanitarian rights[9] and principles[10] encoded in it.

After determining the circumstances under which it is permissible to use lethal force, the subsequent chapter will deal with the issue of where responsibility lies when an autonomous system violates one of the principles of International Humanitarian Law and commits an act that constitutes a war crime. This question is trickier, because amongst critics there seems to be a theoretical assumption that liability for a criminal act is limited to the agent that performs the action (Anderson and Waxman, 2013: 17). However, assigning responsibility according to this assumption is difficult in the case of a machine, as holding it culpable would mean that it is considered to be an artificial intentional agent with full moral capacities. Ascribing personhood to a machine is objectionable to many, so the alternative that we are left with is to assign responsibility to the persons creating or using the machine. There are, however, existing legal mechanisms that can be adapted to assign responsibility, and they do not hinge on the status of the autonomous system as an artificial agent. If these prove inadequate, one could hold either the commanding officer or the programmers liable. Opponents to autonomous weapons argue against this, claiming that as the autonomy of the system becomes more sophisticated, and its actions become further removed from their original protocols, it becomes more difficult to establish causal responsibility to the system's creators and operators (Johnson, 2014: 2). It is important to establish some sort of causal responsibility practices, as artificial intelligence that can make lethal decisions without any human or humans bearing responsibility is disconcerting, for reasons that will be discussed more thoroughly in the chapter on responsibility.

Lastly, I will summarise my discussion, concluding that given the fact that there are no suitable grounds (on neither the objection that their use inherently contravenes International Humanitarian Law nor the objection that there are no suitable candidates for responsibility) for

[9] The two most important pillars of humanitarian law are the right to life and to human dignity
[10] The principle of discrimination and the principle of proportionality will be discussed in chapter 4 on permissible lethal force.

banning the use of autonomous weapons systems outright, the regulation of their use is the best and only viable option. Though such regulation may be a time consuming process, there are many existing legal practices discussed in this thesis that can be used as rough measures of arms control in the interterm. Nonetheless, there are still several topics that need to be addressed in the future, some of which are briefly mentioned here. A topic for future research would be establishing an internationally-accepted definition of autonomous weapons systems that includes the differences between fully autonomous systems and ones with partial autonomy. Furthermore, there is a need to eliminate the ambiguities in the principle of proportionality, which has not been as robustly explained as the other pillar of permissible lethal force, the principle of distinction. Additionally, a fruitful pursuit would be to delineating new criminal offences specifically for the use autonomous weapons systems and for reckless commanding officers, and developing industry standards that manufactures of autonomous weapons systems need to abide by.

The ethics of warfare has been deliberated on throughout human history, and the debate is always revitalised with the development of new weapons, from the long bow through to gunpowder (O'Connell, 2014: 224). The outcome of these ethical discourses inevitably lead to regulations formalized in legislation, thus the overall aim in this dissertation is to explore the ethical arguments around autonomous weapons systems in order to guide future legal legislation, specifically with regard to arms control regulation. There are two possible outcomes to such an investigation: either the existing legal prohibitions will be found to be inadequate and need to be replaced, or they will be found to be sufficient and applied to the new weapon (Anderson and Waxman, 2013: 9). It is my position that while there are existing mechanisms that could potentially be applied to autonomous weapons, there are also other areas that are inadequate and that either need to be revised or replaced with new laws in order to accommodate the increasing use of autonomous weapons systems. Autonomous weapons are here to stay, and therefore we need to address the issues surrounding their use in order to ensure that they are adequately regulated.

2. Artificial Intelligence

The automation of robotics systems[11] and the creation of artificial intelligence[12] (AI) are two of the most ambitious and long-standing goals of modern computational science. While the two are different, automation being more practical and AI being more theoretically inclined, they are closely linked and indeed, the idea of whether or not a computer could be intelligent stemmed from the question of automating tasks.[13] Any robotics system that performs automated tasks would presumably have at least a crude level of AI, enough to reasonably determine what choices to make in light of the task it is trying to or, "desires" to, accomplish. Any machine with AI would necessarily have the ability to perform tasks without human supervision. While there are various levels of autonomy in weapons systems, most would agree that a fully autonomous weapon system would need some form of AI to qualify as fully autonomous. Thus when defining autonomous weapons systems, it is necessary to understand basic AI theory, which is the subject matter of this chapter.

Artificial Intelligence as a field can be defined as the study of how to build and/or program computers to enable them to function in a manner similar to the human mind (Boden, 1990: 1). There are disagreements about whether or not it is, in fact, possible to build a machine with human-like intelligence. These arguments are important to understand because they are the same arguments that people use to claim or deny that an autonomous weapon could or could not sufficiently match a human agent's decision-making process to comply with international law. The two best-known opposing views on the possibility of artificial intelligence belong to Alan Turing and John Searle. Each of the perspectives will be discussed shortly, but before we discuss whether or not artificial intelligence is possible I will give an overview of the

[11] Robot automation is concerned with creating machines that can perform certain tasks with a high degree of autonomy, typically while receiving inputs from the environment to guide their actions, without human supervision.

[12] This concept will be more clearly defined in this chapter, but a rudimentary starting point is to think of it as human-like intelligence in a machine.

[13] This is elaborated on in the discussion of Alan Turing. Turing was prompted to ask whether machines could think after his efforts to formalise a general model of calculation and computation.

philosophical idea of intelligence, since without an understanding of what exactly we are trying to replicate, we would have no idea whether the goals of artificial intelligence could be reached.

2.1. Intelligence

Many AI enthusiasts consider the field as the science of intelligence in general, believing that its goal is to provide a systematic theory that can explain the general categories of intelligence with the goal of replicating them (Boden, 1990: 1). This is an extremely broad definition that is not useful to understanding the AI debate at all. As Simon Blackburn points out, computers are already able to do incredible things that surpass normal human ability. A case in point: we would certainly say someone in a chair working out the next decimal of *pi* is thinking, thus we ought to (for the sake of logical consistency) attribute the same characteristic to a computer doing the same thing (Blackburn, 2009: 85). There are an infinite number of things the human mind can do, and in some instances there are machines capable of performing such activities[14], perhaps even better than humans, yet not even all supporters of AI would count these machines as intelligent, so the standard of intelligence needs further refinement.

Before one can meaningfully ask whether or not machines are intelligent, one has to define exactly what is meant by intelligence, and this definition is not as obvious as one might think. A good place to start is to consider the everyday use of the word "intelligence". Fred Dretske discusses the manner in which we use "intelligence", stating that when one applies the term, it is normally used in two ways (Dretske, 1993: 201). The first use is as an attribute everyone has, but that some have more than others, such that "intelligence" is meant in a comparative manner, for example, "this guppy is more intelligent than the average guppy" (*ibid*.). The second manner is as a certain set of characteristics that are *enough* to qualify as constituting intelligence (*ibid*.).

If we mean intelligence in a comparative manner, then we would merely be saying that a particular machine is intelligent compared to others of its kind, which would mean all

[14] For example, mathematical calculations and deductive problem solving.

machines have some level of intelligence (*ibid*.). While it is certainly possible to say that a particular machine (or in the case of this dissertation, a particular weapon) is more "intelligent" than another of its kind, a comparative understanding of intelligence does not help us to understand AI, and it certainly does not further the debate on whether machine "intelligence" is equivalent to human "intelligence". Therefore, in the study of artificial intelligence, when we speak of machine intelligence, we must be referring to a certain minimum requirement for intelligence (*ibid*.). We need a clearer picture of what this minimum could be, not only to guide those that engineer and develop AI, but also to guide those that will interact with these systems and to determine whether or not we should treat machines with AI like we would treat people. In any introduction to AI and theories of intelligence, the two starting arguments are those of Alan Turing and his most vocal opponent John Searle. Each of them had their own conceptions of what a minimum requirement for intelligence was.

2.2. Alan Turing and Computing Intelligence

Alan Turing was a well known computer scientist and mathematician who, to this day, is still mentioned in discussions on philosophy of mind and artificial intelligence (Hodges, 2013). Turing formalised many concepts in the field of computer science, such as algorithms[15] and computation[16], and is considered to be the father of theoretical computer science and of artificial intelligence (*ibid*.). Any discussion of artificial intelligence would not be complete without examining Turing and his work.

Turing was initially interested in one of the foundational questions of computer science, specifically whether or not certain tasks where computable and if they could be completed by following a specific set of instructions and adhering to predetermined rules (Barker-Plumber, 2016). In order to determine if this was possible, he had to come up with a general notion of how computation or algorithms work, as opposed to a specific instance thereof. This caused him to develop the idea of a Turing Machine. A Turing Machine is a mathematical model of a hypothetical computing machine, or simply a state machine[17], that is ruled by a predefined set

[15] The process or set of rules adhered to in calculation or problem solving, essential in computing.
[16] Equivalent to calculation.
[17] A machine that has a finite number of states or stable conditions that it can be in at any given time.

of instructions that govern how that machine moves between states in order to determine a result from a set of variables (*ibid*.). It is defined by a mathematical model of computation.

In Turing's supposition, he presented an example of a machine that has an infinite one dimensional strip of tape, which is divided into cells, and each cell stores the inputs or outputs of computation (in the forms of 1s and 0s) that codes up to a solution or answer (*ibid*.). At any particular time, the machine is in a particular state. The machine is presented with a task or problem, after which it reads the cells one at a time. Given the state that it is in and the cell that it is reading at that time, it will perform a certain task as determined by a log book[18], which transforms the tape into a new string of 1s and 0s, before moving onto the next state or task (*ibid*.). If the machine reaches a halting state, that is to say, its task is complete, what is left of the string or tape is the answer or solution to the original task (*ibid*.). When a machine can do this (reach a halting state), it is called Turing complete. This is a very simple model, but it is the essence of computation and it essentially served as a blueprint for modern digital computers. The fundamental function of a Turing machine was calculation[19], that is, a type of mathematical determination. For Turing, the idea of Turing machines, computation and calculation gave rise to the idea of machine intelligence.

Turing was a vocal advocate for the possibility of "thinking" machines, producing one of the most influential papers on the subject, "Computing Machinery and Intelligence" in 1950. Turing limits his discussion of machine intelligence to digital computers[20] (1950: 436), since digital computers have three components, which he believed are similar to the structures and functions of the human mind, namely the store[21], the executive unit[22], and the control[23] (1950: 437). This analogy is the basis of contemporary functionalist[24] theories of mind, which draw from Hobbes' conception that the mind is a "calculating machine" and views the mind as a

[18] A rule book with a very simple set of instructions.
[19] Calculation is equivalent to computation.
[20] A machine capable of calculation and problem solving.
[21] Where information is deposited, akin to memory.
[22] Performs individual operations and calculations.
[23] Directs the operations performed by the executive unit and ensures that rules are followed correctly.
[24] Functionalism is the theory that mental states are defined by what their function is rather than by what they are made of. According to functionalism, computation and intelligence are essentially the management of (uninterpreted) symbols according to prescribes rules (Boden, 1990: 4)

computer, and intelligence as a computation (Levin, 2016). Turing took the idea of calculation in machines as analogous to the idea of how calculation works in the human mind. He saw the human mind as a computation that arises from the brain, which acts as a computing machine (*ibid*.).[25] Like a computer, according to Turing, the mind takes inputs form the external world, which gives rise to a particular mental state, and has a certain output, namely a physical state.

The idea that consciousness[26] is determinable based on computation stems from the intuition that since it is not possible to have direct access to another person's mental states, the only way we can infer that they have mental states is based on their behaviour (*ibid*.). For example, if they act afraid, then we must believe they are afraid, regardless of whether the associated physical or biological components of fear (like adrenaline) are present. We can only make inferences based on the products or functions that are produced by a mental state, such that if something appears to have a particular mental state we must assume it does indeed have that mental state (*ibid*.). For functionalists, mental states are "realised", and the same mental state can be realised by different people based on different physical states (*ibid*.). Turing believed a digital computer can perform calculations the way a human mind can, and indeed human calculation works in the same manner as a digital computer (Turing, 1950: 436). This is why Turing believed that a machine that could calculate or compute the way that the mind does and that this was sufficient for intelligence.

In his flagship paper, Turing sought to answer the question of whether or not machines could think. Since that question is too complex (or too "meaningless", in Turing's words) to answer satisfactorily, Turing instead asked if a computer could cause us to believe that it is thinking, say by playing an imitation game (Dowe et Oppy, 2016). Here he sets out the so-called Turing test, where intelligence would be ascribed by enquiring whether the machine could play an imitation game (Turing, 1950: 433). Essentially, the Turing test is as follows: suppose there is a digital computer and a human behind a screen, both being interrogated separately with the same questions (*ibid*. 433-4). If the interrogator cannot differentiate the man

[25] This is not strictly a digital computer, but rather refers to a more general machine that can manipulate symbols, like a Turing machine.
[26] Roughly referring to internal mental states.

and the machine, then the machine must be credited with the same intelligence one would accord to the human (*ibid*. 434). This test is fundamentally functionalist in its assumption that intelligence is, in general, explicable in terms of effective procedures implemented in the brain, thus Turing argued that intelligence could be simulated by an expert machine, which became known as a Turing machine (Boden, 1990: 4). If intelligence is explicable in terms of the ability to perform certain functions, as Turing believed, a machine that could perform similar functions should be considered intelligent. As mentioned, this functional equivalence was deemed a satisfactory measure for Turing, on the basis that we have no direct access to the mental states of other people.

Turing discusses some possible objections to a thinking machine, but one that is relevant to our discussion is that while some critics may allow that a machine could perform the kinds of activities that Turing believes they can indistinguishably from a human, they believe that a machine will lack further human qualities, such as having emotional or ethical capacities and will thus be unable to perform well enough to be attributed social intelligence (Turing, 1950: 447).[27] Turing says there is no support for such arguments, and that they are possibly based on induction[28], since no one has yet encountered a computer who could implement one of these subtle capabilities (*ibid*.). Turing also asserted that much of the emotional and social capabilities that humans have are not *a priori* but are learnt, thus it could be possible for a machine to learn them. Turing also discusses Lady Lovelace's objection, which is still a popular protest, that a machine will only have the capabilities that we endow it with (*ibid*. 450). Turing believed that this was irrelevant, since one day computers will be able to learn for themselves (*ibid*.), which they can today.[29]

For Turing, a Turing machine that could play his imitation game and pass the Turing test was enough for the ascription of intelligence. In other words, intelligent behaviour that is

[27] The ability to understand and manage social relationships, like friendship, family relationships, romance, part of which is self-awareness and awareness of other's "self".
[28] Inductive reasoning generalises multiple instances in strong support of a conclusion, mistaking that the conclusion is probable for it being certain. In this case, the inductive conclusion would be that just because such a machine has never existed, it is not possible for it to exist. This argument is fallacious.
[29] For example, IBM's Watson has the ability to learn from the information presented to it, recognising patterns and drawing new conclusions that it did not have from the beginning.

functionally similar to the intelligent behaviour exhibited by humans is enough to qualify as intelligence. In terms of autonomous weapons systems, a Turing functionalist would argue that a system of this kind that performs similarly to a human directing a non-autonomous system should be considered to have human-level intelligence and to be capable of making the kinds of decisions that we trust humans to make. This is a bold conclusion to accept about autonomous weapons systems, because ascribing intelligence to an autonomous weapon is a life or death matter; if we accept Turing's functionalism but it is erroneous, weaponised Turing AI will have the capacity to take lives without understanding why they do so and the collateral damage will likely be high. Thus we have to be exceedingly certain that Turing's test is a satisfactory gauge for endowing systems with lethal capacities.

A weakness of the Turing test is that the computer is essentially trying to trick the interrogator into thinking that it is human in a once off meeting, which is a flimsy basis for declaring a machine to possess human-like intelligence. The few machines that have passed the Turing test have done so through means that some find less than honest. The first machine to pass did so in 1966 and was called ELIZA, who passed by simulating a psychologist, specifically by reflecting the interrogators questions back to them and making them talk about themselves as opposed to interrogating her (Weizenbaum, 1966: 42). The latest case was in 2014 when the machine dubbed EUGENE GOOSTMAN posed as a 13 year old Ukrainian boy (McCoy, 2014). EUGENE was able to dismiss its English language mistakes since English was not the "boy's" first language. In both these cases the designer employed a strategy to fool the integrators, as opposed to the machine itself actually playing the imitation game successfully. This does not, however, mean that the "smart" machines we have today are any less impressive in terms of their sophistication and AI. Despite the fact that they may not be able to fool us into believing that they are human in a conversation, they can functionally perform the specific task that they were designed for as well as, if not better than, a human performing the same task. When one looks at autonomous weapons, there are weapons that can track enemy combatants far better than humans, through radar and other robotic sensor technology[30], and can make decisions to engage targets at least as consistently as human soldiers do. We must consider these machines to be functionally intelligent, even though we would not converse with them

[30] This point is discussed further in Chapter 3 on Autonomous Weapons Systems, under the subsection of Potential Benefits.

like we would with a human. These machines may currently lack the emotional and social capacities that are necessary for them to be allowed full autonomy in lethal decision making without human supervision, as required under International Humanitarian Law, but that does not mean that this will always be the case. Turing's idea of machine intelligence should not be wholly dismissed, and I will reprise the test in light of autonomous weapons in a later subsection.[31]

Before discussing Searle's refutation of the Turing test, there is a noteworthy criticism offered by Bruce Edmonds that is relevant in light of the typical use that is made of autonomous weapons. Edmonds accuses Turing of being "cunning" in his formulation of the Turing test. Here, intelligence is not dependent on conceptual characteristics, but rather it hinges on the ability of the machine to perform to a certain standard, specifically, its ability to fulfil a social role in a once-off meeting, and he criticises Turing for not setting a minimum duration for his test to run (Edmonds, 2000: 420). Edmonds states that part of our perception of intelligence is the ability to "out-think" one another, especially over a period of time (*ibid*.). He writes that it would be easy for a machine to trick us in a once-off meeting, but over time, given that our knowledge is situated temporally, it would become obvious that the machine didn't have a history, a personality, memories, experience, etc. (*ibid*. 419-20). He proposes a Long Term Turing Test (LTTT) in order to make the Turing Test more meaningful (*ibid*. 420). A LTTT would be more difficult to pass, and so a machine able to pass it would possess a more genuine form of artificial intelligence (*ibid*.). Edmonds, in fact, later questions how "artificial" such a machine would be; if the machine had been constructed, then set in an environment and had learnt from the environment and had the ability to alter its own original store of information, there is no way to tell how much, if any, of the original store remains (*ibid*. 422). As such, we could not say that such intelligence is "artificial" in the sense that it is merely an imitation, since the computer created its "self". According to Edmonds, it would be considered to be no different to a human, meaning that it qualifies for full personhood. This kind of a machine would have the same rights and duties as a natural moral agent.

[31] Subsection 2.5 entitled "Reprising the Turing Test".

While some autonomous weapons could be deployed for a specific mission only, some will be deployed over the long term, and given that they have the ability to learn from the environment they are deployed in, and possibly the ability to alter the original store of information it was endowed with[32], Edmond's LTTT is relevant. In testing for intelligence, the ability of the autonomous system to play the imitation game would have to be tested over the long term, because in the case of a weapon with lethal capabilities, it is better to be over cautious and thorough with the testing process. The LTTT has a curious consequence though, as according to the LTTT, a weapon that could behave in a manner that is indistinguishable from human behaviour over the long-term could be attributed full personhood.[33] The personhood of an autonomous weapon is critical to the question as to whether or not they could be held liable for war crimes, but I will elaborate this point later in the chapter on responsibility.[34]

Turing's test may attribute intelligence too readily for most cognitive scientists and AI theorists, but there is a valuable point that we can take away from it: we have no access to the mental states of other persons, so while behaviour might not be enough to actually measure the presence of intelligence, it does guide us in our actions and in how to interact with a machine that is at least behaviourally intelligent. I will return to this point in my later reprisal of the Turing test, but first I want to enquire, if behaviour is not enough to indicate the presence of intelligence, then what is? One of Turing's greatest critics, John Searle, believed that the necessary condition for intelligence is intentionality. Searle's criticises Turing for the mistake he made of equating behaviour to intelligence, and his criticism relies heavily on the concept of intentionality. Searle's argument and the concept of intentionality is critical to understand as a foundation for any discussion on holding a machine liable for its actions.

2.3. John Searle and the Chinese Room

Many critics believe that Turing too eagerly attributed machines with human intelligence in the Turing test (Boden, 1990: 4-6). A popular method of refuting Turing's

[32] A full discussion of autonomous weapons and their capabilities will be made in Chapter three.
[33] The status of being a person, including all associated rights and duties.
[34] Chapter five

position is to use some kind of anti-behaviourist argument to dismiss the imitation game as a successful criterion of intelligence (*ibid*. 4). The most successful of these types of arguments need only show that the computers are not necessarily intelligent as based on behaviour (*ibid*.). In other words, the presence of seemingly intelligent behaviour does not necessarily imply the presence of genuine intelligence. Most AI theorists would agree that intelligence necessarily involves causal processes (computations) of a certain systematic sort and that behaviour, no matter how uncanny, is not enough in itself to qualify its bearer as intelligent (*ibid*. 5). The anti-behaviourist argument boils down to the idea that such a demonstration of intelligence could merely be a simulation and does not necessarily count as true intelligence. The most notable critic in this school is John Searle. In most discussions on artificial intelligence, Searle's position is mentioned shortly after Turing's.

Searle presented his famous refutation of Turing and functionalism in his paper titled "Minds, Brains and Programs" (*ibid*.). Initially he distinguishes between "strong AI" and "weak AI" (Searle, 1980: 417). Weak AI views artificial intelligence as a tool that enables us to model and understand intelligence better without claiming to actually replicate it, as strong AI does (*ibid*.). Searle is stalwartly opposed to the claims of strong AI and Turing's assertion that a machine behaving similarly enough to human intelligence could be considered genuinely intelligent, because he believes that devotees of strong AI and Turing machines not only believe that the machine performs the same calculations as the human mind, but also that it understands questions put to it and provides its own answers, and that this literally replicates human ability (*ibid*.). Searle uses a thought experiment, known as the Chinese room, in order to show how this could never be the case.

The experiment is as follows; suppose Searle is in a room with a batch of Chinese writing and he has no knowledge of Chinese, such that the writing is merely a series of various "squiggles" to him (*ibid*. 417-8). Furthermore, suppose there is a second batch of Chinese writing, which appears to him as a series of "squoggles", along with an English rule-book to facilitate the correlation of the first batch with the second (*ibid*. 418). This is meant to be representative of the way a digital computer processes information, and is effectively what a Turing machine does. If one considers how a computer (or a Turing machine) works, it takes

inputs and provides outputs according to fixed rules; to effect a computation it has to turn any data into strings of 0s and 1s, or electrical patterns (Blackburn, 2009: 87). This is something like what Searle is doing in the Chinese room.

Now suppose that in stead of starting with two batches on either side of him, the first batch of "squiggles" are being fed into the room, and Searle then uses the English rule-book to feed out the correct corresponding "squoggles" of the second batch (Searle, 1980: 417). Unknown to Searle, the "squiggles" are actually questions from Chinese-speakers, and he is unwittingly providing the correct answers in Chinese (*ibid*.). Searle has no understanding of Chinese or of what is transpiring (*ibid*.). He argues that this is essentially what a Turing machine is doing, namely processing inputs to outputs according to fixed rules without understanding what they are or why (*ibid*.). As he puts it, Turing machines still lack intentionality (*ibid*.).[35]

Searle was trying to show that a Turing machine would merely be mimicking human intelligence but would not actually have genuine intelligence. Searle believed that a Turing machine would be helpful for understanding the nature of intelligence, in line with weak AI, but he rejected the idea that a Turing machine would literally replicate human intelligence like the claims of strong AI. For Searle, the minimum requirement for genuine intelligence would be intentionality; and machine, lacking the neuroprotein brain that gives rise to our own intentionality and consciousness, would necessarily lack intentionality and therefore could never be considered to be intelligent in the capacity that humans are. While most AI theorists agree that behaviour on its own is not enough, I do not find Searle's position entirely convincing, for similar reasons that Margret Boden and Tracey Henley criticise him.

The first worthy criticism of Searle's argument which I find a convincing refutation is offered by Margaret Boden. She opposes two claims of Searle's argument; (i) that functionalist

[35] Intentionality here is meant in sense that it is used in philosophy of mind. Its definition are made clear later, but broadly, it refers to the internal mental property of having awareness or understanding regarding external properties. It is different to the way the term "intention" is used in the law, which will be referred to in later chapters.

models are theoretical and do not explain the practical process of cognition, and (ii) computer hardware, unlike neuroprotein, lacks the right "stuff" for intelligence (Boden, 1990: 89). Searle's first claim is that computers merely perform symbol manipulation based on instantiating processes and syntactic rules, and therefore lack any understanding of the symbols themselves (*ibid*.). Boden replies that Searle's description involves a "category-mistake" misidentifying the brain as the source of cognition and intelligence itself, as opposed to it being the causal basis of intelligence (*ibid*. 96). Essentially she concedes that while the brain does seem to be the substrate that supports our consciousness or intentionality, it does not mean that computer hardware is unable to provide the necessary substrate for a machine. This is the assumption that Searle's argument rests on. Searle's second claim is that intelligence (and intentionality) is a biological phenomenon (*ibid*. 91). Boden states that Searle provides no basis for his claim and furthermore has no proof of this, and that many theorists hold that intentionality could be a psychological function or a logical function (*ibid*. 92-3).

Tracey Henley, a neuro-behaviourist, who echoes Boden's argument, articulates the second objection to Searle. Henley believes that on reflection, it does seem that Searle makes bold claims with his argument, since Searle trying to prove that AI is *a priori* not possible (1990: 45-6). Henley surmises Searle's argument as follows:

(i) Humans have intentionality,
(ii) Intentionality is the result of the causal powers of neuroprotein brains,
(iii) Computers do not have neuroprotein brains; therefore,
(iv) Computer can never have intentionality
(Henley, 1990: 46-7)

Henley criticises Searle's claim (that only humans have these causal powers) because this position is not testable; Searle makes a series of claims without support or proof that one could not disprove (*ibid*. 47). In light of his unfalsifiable argument, Henley calls Searle a debunker who make bold claims that intentionality is linked to having a brain, and a machine could not have it, but there is no convincing argument to show that this is the case (*ibid*. 46).

Dretske provides an argument to show that behaviour itself is not a satisfactory qualifier for intelligence that I find more convincing than Searle's, and which does not raise the issues Boden and Henley found in Searle's argument. Dretske successfully demonstrated that while behaviour can appear to be intelligent, there may not be any active intelligent reasoning behind it and that intelligent behaviour has to be accompanied by understanding and governed by thinking (1993: 202). That is to say, cognitive representations have to be related to behaviour (Dretske, 1993: 203). To illustrate his point, take the example of a zebra running from a lion. The zebra has a cognitive representation of a lion, which causes the zebra to behave in a certain way when it recognises the concept "lion" (*ibid*.). The zebra is behaving in a manner that we can ascribe reason to, but it is doubtful whether the zebra has any more cognition behind the action than basic survival instinct (*ibid*.). The zebra lacks the necessary understanding or reasoning abilities for its act to qualify as intentional and thereby intelligent. This leads to the conclusion that there is an attribute of intelligence that goes beyond behaviour, because instinctual automatic responses cannot meaningfully be described as intelligent (*ibid*. 204). Dretske's qualifies the minimum requirement for "intelligence" or "thinking" as having some kind of cognitive representation, or rather, intelligent behaviour is constituted by understanding and reason (*ibid*. 203).

Dretske expands on this position, writing that often we believe thinking, in the sense of mere mental processings, is not enough on its own, because then any activity that happens in the brain would constitute intelligence, even automatic responses. Dretske writes that this does not seem an adequate criterion to attribute intelligence by, for we would not attribute a zebra with the same intelligence as a human has, since the zebra seems to merely be reacting instinctively rather than understanding *why* it is doing so beyond recognising some danger. Essentially Dretske argues that even though the zebra's behaviour conforms to its thinking, it is not explained thoroughly by the thinking (*ibid*. 204). So automatic behaviour, even if there's a reason for it, isn't enough since behaviour alone is not a manifestation of intelligence (*ibid*. 206). Dretske concludes that thoughts must be linked to the behaviour through reason (*ibid*. 207).

I agree with Boden and Henley; Searle does seem to make some bold claims that rest on yet-to-be falsified assumptions. I have already conceded that mere behaviour may not a satisfactory indicator of the presence of genuine intelligence, at least not enough for us to endow a machine with personhood and full moral agency. I do agree with that intentionality appears to be a better qualifier; however, I do not believe that Searle made a satisfactorily convincing argument to prove that intentionality is something that a machine cannot *de facto* possess. Considering all these arguments, I agree with Dretske's position that at minimum intelligence would require thinking in a specific way, namely having a reason to behave in a certain way and understanding this reason. Thus, there is something more than behavioural equivalence to the ascription of genuine intelligence and this is likely something akin to understanding the reason behind behaviour, which could also be described by beliefs and desires. Intelligent behaviour is necessarily the causal result of mental process or understanding the reason for behaviour. Thus if a machine were to processes information in a manner linking the processing of it to its behaviour, and have a reason to do so, accompanied by awareness of the reason, it could be considered to have human-like intelligence. If an autonomous weapon had this level of intelligence, it would qualify for full personhood, and could therefore be held liable for its actions. In light of the bearing intentionality has on responsibility, I believe it to be worthy of closer examination.

2.4. Intentionality

Searle believed that intentionality is what at minimum constitutes intelligence and he argued that machines, lacking the "hardware" we have, could never be considered genuinely intelligent. He, like other critics of the functionalist and behaviourist philosophies of intelligence, argues against the idea that mere behaviour based on instantiating processes is a sufficient basis for intelligence (1980: 422). Searle's Chinese Room argument may be convincing to some, but I do not find his argument overwhelmingly convincing, for the criticisms offered by Boden and Henley. I believe that Dretske provides a more compelling reason to look for a characteristic over and above behaviour, thus I concede that Turing's functionalist view leaves something to be desired, and that something is likely intentionality. It is therefore prudent to understand what intentionality is and to establish what could count as intentionality in machines, as intentionality is something that we need for personhood and full

moral agency. Without it, we would be wholly unable to assign responsibility to the machine itself for its actions.[36]

Intentionality is a phenomenological[37] concept that stems from a dualist[38] perspective. It arises from the acknowledgement that minds seems to have properties that are not explained or contained in the psychical world (Jacob, 2014). The concept of intentionality was first popularised in philosophy of mind due to the efforts of Franz Brentano, a German philosopher and psychologist who was most famous for his work on this subject (*ibid*.). Brentano took the old psychological concept of intentionality and introduced it to philosophy of mind, in order to explain the nuances of consciousness (*ibid*.).

The word "intentionality" derives from the Latin *intentio* which means to be directed towards something (*ibid*.). In philosophy of mind, intentionality is a feature of mental states and broadly refers to their being directed towards some property or state. Brentano summarised intentionality as "aboutness", meaning that property of mental states to be "about" or relate to the external world (*ibid*.). In this way, he believed that every intentional state (like a "belief" or "desire") has an intentional object that it is about, such that a mental state like a thought points to or refers to a target (*ibid*.). Brentano also differentiated levels of intentionality. The most basic level of intentionality entails a person's mental states about non-mental, physical things, for example, a person's beliefs about a chair (*ibid*.). Higher-order intentionality[39] denoted a person's mental state regarding the mental states of others, for example, a person's belief about the beliefs of other persons (*ibid*.). Brentano, like other dualists, examined the nature of mental states and how different mental states can be realised from the same physical referent. He further noted how it is possible to have mental states about a referent that does not exist, and concluded that mental states and intentionality is not merely a physical property

[36] This is not to be confused with the idea that machine intentionality is necessary for holing autonomous weapons responsible for their actions, since there are still other viable candidates for responsibility, but this point will be discussed fully in Chapter 5 on Responsibility.
[37] Structures of consciousness experienced form a first-person point of view
[38] The view that the mind and the brain are not identical and that consciousness is not merely a physical phenomena. This is the opposite of the physicalist perspectives, like functionalism and behaviourism, that consciousness can be explained in terms of physical phenomena.
[39] This concept will be discussed again in light of Dennett's argument on machine liability, in Chapter 5 on Responsibility.

(*ibid*.). In other words, functionalism and behaviourism, that explicate consciousness as a physical property, is incorrect on this internalist view.

Searle believes that intentionality is a result of the causal powers of the human brain (Searle: 1980: 422). In other words, intentionality is a product of our neuroprotein human brain and that is what grants them intelligence. He allows that other substrates may support intentionality and therefore intelligence (i.e. intelligence is not only found in the human brain), but maintains that a computer is not such a thing (*ibid*.). Searle asserts that computers, being rule-governed, are limited to syntax.[40] In that way, a Turing machine does not understand the string of tape any more than recognizing 1s and 0s. Contrary to this, human beings can be said to understand and have intelligence based on intentionality. Using Searlean terminology, computers have syntactic intelligence, while humans have semantic[41] intelligence (*ibid*.). This simply means that computers have knowledge of nothing more that the syntax rules that govern the machine, while humans have an understanding of concepts, some kind of cognitive representations and mental states that counts as intentional, and therefore humans are intelligent (*ibid*.). As mentioned, Dreske took intentionality as an understanding or a type of awareness of the reason for behaviour, and Brentano believed that an intentional mental state is one that is directed towards or is about something. Broadly, intentionality would entail some kind of understanding or awareness. Searle believed that this understanding or awareness is a product of the human brain.

Intentionalists replies to Turing (like the those of Searle and Dretske) normally use arguments to show that machines could not be intelligent, because they lack intentionality (Boden, 1990: 5). Some, like Searle, assert that mental states are contingent on the physical states and that it is only possible to share mental states as far as corresponding physical states are shared, and in the case of intentionality, it corresponds to the physical state of the neuroprotein brain. Searle's assertion is supported by the following example: when a person suffers significant damage to their brains, specifically to the frontal lobe, they usually

[40] A linguistic terms referring to he arrangement or words or phrases according to rules.
[41] The understanding of the meaning behind words or phrases beyond understanding the rules to arrange them by.

experience changes in their personality and in they way that they experience and perceive the physical world. Intentionalist like Searle even go as far as to say even if the machine performed as Turing had envisioned, it would not really be intelligent because no computers could conceivably think or understand; i.e. there is no genuine intelligence without intentionality (*ibid*.). In the end Searle concludes that it may be possible to build a computer that can simulate the "computations" that the human mind performs but it would not be truly intelligent (weak AI); it would not understand and it would lack the intentionality required for it to be considered genuinely intelligent (strong AI) (Searle: 1980: 422).

If we accept that behaviour is not a sufficient indicator of the presence of genuine intelligence, and I believe that there is a strong case for this, then there must be something more. Given the arguments put forth by Dretske and Brentano, it is likely that this something is intentionality. However, I think it is more useful to think of intentionality in a broader sense as understanding or awareness and not in the "human-exclusive" sense that Searle meant it. But even if we accept intentionality as the minimum requirement for intelligence, we still do not know where it comes from, or indeed if other people have it. After all, my ascription of consciousness to other human beings is based on assumptions I make, assumptions inferred from their behaviour. Therefore I do not find that intentionalist arguments discredits the functionalist thesis completely, since the functionalists' foundational claim[42] is still true; behaviour is the sole source from which one can infer the presence of intelligence, or intentionality. This may not be satisfactory to indicate the presence of genuine intelligence, but it does guide our own actions and understanding of a machine capable of playing Turing's imitation game, and therefore it is a useful position. While Turing's assertion was found too broad, there are merits in his view that are worth re-examining, so now I will discuss a reformulation of the Turing Test in light of autonomous weapons systems.

[42] The claim that we have no direct access to another person's mental states and thus we can only assume that they have mental states similar to our own based on their behaviour and the functionality of those mental states.

2.5. Reprising the Turing Test

Various proposals for evaluating whether or not artificial entities are rational agents exist, and the Turing test is only one such an example, albeit the most basic and introductory one in AI theory. As discussed, the test is philosophically controversial because it equates behaviour with mental states, or at least treats sophisticated behaviour as a crude proxy for deeper mental states. While it is one thing to treat complex behaviour as a proxy for cognition, it is another to treat it as a *reliable* proxy for consciousness (Ohlin, 2016: 13). The difference lies in the latter claim that behaviour that is seemingly intelligent must be enough to qualify as genuine intelligence (like with Searle's strong AI). The first understands that we may never be able to point to "intelligence" in itself, and that it is more a useful concept for us to employ when dealing with "thinking" machines.

Jens Ohlin takes the Turing test and modifies it to reach a less controversial conclusion, where the artificial agent would qualify for personhood if its behaviour *simpliciter*[43] were virtually indistinguishable from the behaviour of a natural human being. The difference is that if a machine passed the Turing test, it does not necessarily mean that we would attribute it with internal mental states, but rather we would treat it as if it were intelligent in order to interact with it (2016: 14). This is a more pragmatic and plausible version of Turing's test: if the artificial being is indistinguishable from a natural person, we would only be able to interact and understand it as if we thought of it as intelligent, and any question of the actual property of intelligence is left out (*ibid*.).

Ohlin takes this pragmatic conclusion a step further, applying it to autonomous weapons. The idea of what he calls the Turing Test for Combatancy would be that the behaviour of the artificial agent participating in combat would from a distance be functionally indistinguishable from any other combatant engaged in armed conflict. This would mean the autonomous weapon does everything that any other combatant would; it engages enemy targets, attempts to destroy them, attempts to comply with the core demands of the laws of war as best as possible, and presumably prioritizes the protection of its allies over that of enemy civilians. More

[43] This entails the agent's behaviour at face value without further qualification, such as any linguistic elements.

importantly, enemy combatants would be forced to interact with the autonomous weapon as if it were a natural person.

This still raises the same anxiety as the Turing test, namely that the behaviour is a simulation and lacks the intentional states of a natural person. In the context of autonomous weapons in armed conflict, Ohlin believes the deeper questions about intentionality are not important for pragmatic purposes, and I would have to agree. A very crude understanding of the standard for determining whether the use of force is permissible is whether a rational agent in the same situation would feel threatened and respond with force. In light of this, Ohlin argues that the standard for determining if force is reasonable is whether the opposing combatant views the autonomous weapon as functionally indistinguishable from any other combatant, in that sense the enemy combatant is required to attribute beliefs and desires and other intentional states in order to understand the entity and interact with it, as an enemy combatant (*ibid*. 15). This is what Ohlin refers to as the Combatant's Stance – the position where you are forced to interact with enemy AI as if it were a natural person so that you can make sense of it.

With the Combatant's Stance in mind, it is at least possible to imagine an autonomous system whose behaviour was such that it could only be understood as operating under the same constraints as a human soldier and thus being subject to them as well. One would need to engage with the system as if it where an enemy *combatant*, not as if it is an enemy *weapon*. As such an autonomous weapon is pragmatically equivalent to an intentional agent and asking larger questions about intentionality is irrelevant for how we interact with it (*ibid*. 16). One might object that the deeper questions of cognition are important for determining whether the system could be considered to be an artificial agent, and consequently whether the rights and duties of combatants can legitimately be ascribed to it. Furthermore, these deeper questions regarding "intentionality" can have a bearing on whether an autonomous system could be considered a morally culpable agent for a violation of the laws of war. If the system is merely copying behaviour then it is not a moral agent and it would not make sense to consider it to be an artificial agent and to endow it with the rights and responsibilities associated with personhood under International Humanitarian Law. The points raised in objections are true, but they do not affect the pragmatic determination of permissible behaviour. Additionally, even

in situations where agents are not culpable, and are not even subjects of *ex post* criminal punishment, the rest of international law still applies (*ibid*, 17).[44] This means that even if AWS are not considered moral agents or culpable, the doctrines and principles of IHL apply and must be adhered to at every other stage by every other actor.

The important conclusion to draw from Ohlin is that if the autonomous system is virtually indistinguishable at a distance from a natural combatant, then all combatants would be forced to treat the system as a natural combatant, not only to understand it but also to know what counter measures are appropriate. There would be no other practical alternative, and victory over an autonomous system requires adopting the Combatants' Stance (*ibid*. 18). This implies that for the purposes of guiding a human agent's actions, an enemy autonomous weapon is no different from an enemy soldier, and the existing structures of the laws of war are a sufficient guide for human agents. One could even take Ohlin's point further, claiming that if an autonomous weapons system is behaviourally indistinguishable from an enemy combatant, in that it appears to at least make sufficiently similar decisions regarding lethal force as human decision makers, it would be reasonable to endow them with lethal capabilities. In other words, if the behaviour exhibited by the system conforms to the standards set out in international law, there is no compelling reason to deny it the ability to decide to employ lethal force, regardless of intentionality.

According to Ohlin, the fact that the autonomous system is not a responsible agent, for the purposes of criminal law, is irrelevant to determining whether it should be viewed and interpreted with the Combatants' stance (2016: 19). Even if one assumes that the autonomous system is entitled to none of the international law protections that flow from moral agency, it would still be the case that engaging the system in battle and understanding its behaviour would require thinking of it as an enemy combatant pursuing particular objectives in accordance with the Rules of Engagement (ROE) (*ibid*, 19).[45] It would be an enemy combatant *simpliciter*, and any investigation into intentional states would be irrelevant for pragmatic interaction. Ohlin

[44] The application and principles of International Humanitarian Law will be discussed in chapter 4.
[45] The Rules of Engagement refer roughly to the military and international law directives that govern behaviour in combat.

also believes this issue is logically independent of the analysis regarding the commander's liability for deploying the AWS (2016: 19). I agree with Ohlin that the issue of when the use of lethal force is permissible, and culpability for unlawful instances of the use of lethal force, are logically independent, and I will now further this assertion by referencing a distinction that Helen Frowe puts forward.

Helen Frowe, a just war theorist, develops a robust account of self-defence which she believes ought to be "action-guiding", and I believe her argument can be paralleled with Ohlin's (2009: 346). According to Frowe, an action-guiding account must tell the moral agent how to act, or what is permissible. She differentiates between the justness of the harm itself (permission to use lethal force) and the liability to bear that harm (linked to issues of accountability for the infliction of unjustified harm), the first category is subjective and the second is objective (2009: 347-8). Much like Ohlin's Combatant's Stance, Frowe's distinction is meant to guide a moral agent, possibly even an artificial moral agent, in what course of action is permitted, while keeping separate the issues of responsibility.

Frowe clarifies her position with an example. If an agent believes someone (X) means to do them a grievous harm, and if they do not act (or react) this grievous harm will befall them, then they are justified in responding with force (Frowe, 2009: 350). Indeed, this subjective test is a part of making the case for self-defence. Whether or not X actually meant them harm (in other words, the internal mental states of X) is a fact not knowable by the agent at the time they choose to act, and so is not relevant to guiding the agent's actions (*ibid.*). X's intentions, according to Frowe, are only relevant to whether or not X is liable to bear the injury they are met with (*ibid*.).

Frowe's argument serves to emphasise Ohlin's point that there is a difference between knowing what action is permissible, and between the matter of responsibility. The point for my argument is that it is not necessary to establish responsibility for autonomous weapons, or to prove that they have intentionality, in order for them to be legally and ethically utilised in combat. If they were behaviourally indistinguishable from a human soldier, and remembering

that we have no access to the mental states of human soldiers, we could view these weapons with the Combatant's Stance. In this way, a human soldier would know when to engage enemy AI regardless of the internal mental states of the AI. Similarly the AI would be able to recognise behaviour that warrants an aggressive response, such that it knows what action is permitted, without a need to establish responsibility on part of the AI. I believe that this human-like behaviour would be enough to guide the development of weaponized AI, since behaviour is the qualifier for reacting to hostilities. This warrants the deployment of these weapons irrespective of being able to establish causal responsibility to the AI itself.

2.6. Summary

Artificial Intelligence is a critical component in the automation of tasks, including the automation of lethal decision making. It is apparent that some kind of definitional consensus on what counts as intelligence is crucial since without a goal, it is difficult to say what we are working towards. Presumably, the goal would be to have a machine that is at least as intelligent as a human performing the same tasks. While there are many different ideas surrounding what should count as intelligence, it is clear that there must be some minimum requirement. There needs to be clarification on what at minimum constitutes intelligence, such that we now what criteria an autonomous weapon system would have to meet before it is granted the freedom to make lethal decisions.

Alan Turing believed that functional and behavioural equivalence was sufficient. After all, we have no way of confirming whether other people have genuine intelligence, all we have to go on is their behaviour. This prompted him to develop an imitation game, the Turing test, in order to determine whether we could consider a machine intelligent in the same way as we would a human. Turing believed that the kinds of machines that could fool us into believing them to be as intelligent as humans, were not far off in his time. Yet very few, if any, Turing machines exist today, and not for lack of effort. Computer scientists have made numerous attempts to build a machine that could pass Turing's test, but as has been shown, this is more difficult than expected. The formulation of the test is more focused on whether a machine could trick us into thinking it has human like intelligence in a once off meeting, as opposed to actually

trying to simulate it. The Turing test leaves something to be desired. If the test is extended to run over a longer period, it does make the conclusion more believable, but it does not remove the anxiety of equating behaviour to consciousness entirely. Thus most AI theorists reject functionalism, using a range of anti-behaviourist arguments and arguments based on intentionality, predominantly presented by Searle, in an effort to discredit behaviour as an adequate minimum gauge for intelligence. Turing's test is not a suitable measure for intelligence on its own. Most experts in philosophy of mind agree that exhibiting human-like behaviour alone does not imply the presence of genuine intelligence.

Internalism is the natural counter point for functionalism. John Searle, a prominent philosopher in this school, believed that Turing was missing the point. A machine capable of playing Turing's game would be simulating intelligence, but would not actually be replicating it. This is because he believed that the heart of intelligence is intentionality, or as Dretske said, understanding the reason underlying the behaviour, or as Brentano said, the property of mental states to refer to or be "about" something. A popular perspective in academia is that intentionality is the component that makes intelligence more than a mere simulation.

I agree that behaviour is not an adequate measure for intelligence, only in reference to the internal non-physical property that we have no way of measuring in any case other than assuming its existence in other beings that behave closely enough to us. Perhaps intentionality is the crux of intelligence, but I do not agree with Searle's unfalsifiable assumption that machines could never have intentionality. Additionally, just because Turing's argument is not wholly convincing does not mean that we should accept Searle's argument as valid. The functionalist position may not be theoretically sound but they are correct when they claim that there is no way to access another person's internal mental states, or what Searle refers to as intentionality. Essentially, if a machine is functionally equivalent then we would have to at least treat it like it is intelligent. The actual equivalence doesn't matter.

Furthermore, I do not believe that it is necessary to undeniably prove the intentional capacity of a machine, at least not for pragmatic purposes. As Ohlin stated, the Turing test

could be reformulated for autonomous weapons to show that intentionality does not affect how we would interact with a system that behaves similarly enough to a human combatant; it only affects whether we could meaningfully assign blame to that system. If an artificial agent that behaved indistinguishably from a human combatant (over the long term, as Edmonds suggested), we would have no choice but to act as if it we an enemy human combatant whatever our position on genuine artificial intelligence might be. Our pragmatic view of autonomous weapons is a separate issue from whether we would ascribe internal mental states to the machine and accordingly, from how we would punish it. This separation is amplified with Frowe's division between the subjective factors that permit a course of action for an agent (the justness of inflicting harm) and the objective factors that determine whether there is liability or blame for that action (the justness of the harm itself). I agree that there is a difference between knowing how to act, or what is permitted, and responsibility. Frowe's division can be used to determine what actions are justified, both for human agents and artificial ones. For AWS, behaviour is good enough to guide the way we act with them (i.e. Ohlin's Combatant Stance) but not enough to ascribe full moral agency or responsibility. Intentionality is required for full moral agency, but I will return to this point in Chapter 5. For now, given the understanding of artificial intelligence furnished in the chapter, I can offer a more meaningful discourse on autonomous weapons systems.

3. Autonomous Weapons Systems

Before delving into the technical discussion about the lawful and ethical use of autonomous weapons systems, we need to first understand exactly what the term "autonomous weapon system" refers to. AWS is an umbrella term referring to a broad scope of weapons systems. There is enduring confusion regarding the distinction between a system that is categorized as automated and one that is deemed to be wholly autonomous[46], and additional misconceptions concerning issues surrounding intentionality, agency, and responsibility for actions in both of the aforementioned types of systems. A broad definition of what AWS entail is useful to ensure in future that legislation caters for technological developments. A definition that caters for future developments is important for the following reason: as AWS become more sophisticated and complex, arms control measures still apply despite technical differences between older and newer models. However, a definition that is not sufficiently narrow will fail to effectively differentiate between a highly automated weapon and an autonomous one. Thus in this chapter, I will attempt to flesh out the essential characteristics that constitutes an fully autonomous AWS, as opposed to one that is merely passively autonomous. The distinction between automated and autonomous systems is important to maintain, since in the former case responsibility for misuse or a violation of international law always lies with the human operator, and in the latter an argument can be made that responsibility lies with the machine itself.[47]

As mentioned, the goal of this chapter is to hone in on a definition of what constitutes a fully autonomous weapon system as opposed to an automated one. This will be done be examining a few viewpoints on these weapons, specifically the views held by the United States of America's Department of Defense (DoD), the International Committee of the Red Cross (ICRC) and Human Rights Watch (HRW). These organizations are the most relevant in this endeavour – the US DoD because they are the furthest with developing AWS (Allen and Wallach, 2013: 125), the ICRC because the Geneva Conventions and the Additional Protocols

[46] There is even more confusion between what constitutes a fully autonomous system as opposed to a semi- or passive autonomous ones.

[47] I will present this argument in Chapter Five, which is on responsibility.

give them a mandate to protect civilians in armed conflict, where these weapons will in all likelihood be used (International Committee of the Red Cross, 2016), and HRW because they are a humanitarian advocacy group that liaises with governments and the United Nations, which makes their considerations relevant in policy-making (Human Rights Watch, 2016). I will start by examining the definitions offered by these groups and distilling them into one that captures the essential functions of AWS that differentiate them from automated weapons. Establishing definitional consensus is important because it affects arms control regulation; in order to employ effective arms control, we need to be clear on precisely what we are regulating. The definition that I establish here will be held throughout the rest of the dissertation, where the aim is to provide a guideline for arms control. Secondly, I will also examine the reasons for the automation of weapons, as well as the principal criticisms AWS face. Taking note of the benefits versus harms associated with these weapons is important; we want to be mindful of what the goals of AWS are when we are regulating them, as well as being mindful of potential shortcomings of AWS in order to compensate for them. Lastly, I will also examine an implication of increasing weapons system autonomy, namely that it removes human agents from the decision to employ lethal force, and the principal criticisms that this implication gives rise to. This issue is a major part of the AWS debate, since the lawful application of lethal force requires human decision making under international law. This is the reason for the emerging of the principle of having "meaningful human control[48]" over weapons system and the decision to apply lethal force, and this norm will also be examined.

3.1. Defining Autonomy in Weapons Systems

AWS is a broad term, and the many organisations that utilise this term have similar perspectives on its definition. However, no universal definition has yet been formalised, so I will discuss a few of the most salient views in order to offer a definition that captures the most essential characteristics of AWS. It is important to formalise a universally applicable definition for the purposes of codifying meaningful arms control measures.

[48] Broadly, this principle articulates the need to have a human supervising AWS. This concept will be more precisely defined later in the chapter.

In 2012, the US DoD issued Directive 3000.09 regulating the use and development of AWS, the aim of which was to set policies for the responsible use and development of weaponised autonomous systems. The document included some definitions of AWS-related terms, and since the US is the leader in AWS technology, it is an appropriate starting point. The goal of this directive was to minimize the probability and consequences of weapons system failure [49] with regard to AWS and to ensure that these weapons function as anticipated (Department of Defense, 2012: 1-2). The DoD defines an AWS as a weapons system that "once activated, can select and engage targets without further intervention by a human operator" (14). Their definition includes AWS whose decisions can be overridden by humans, since the original decision to attack remains with the system (*ibid.*). While there are various levels of involvement between humans and AWS[50], only the levels where a program can select a target and initiate an attack without any human input are instances where the system can be considered to be fully autonomous (O'Connell, 2014: 226). Any situation where a human must first approve the attack effectively means the system is semi-autonomous or passively autonomous, and the current laws relating to war and responsibility applies to the operator (*ibid*. 227).

The ICRC, who protects civilians in armed conflict, defines AWS as weapons that have full autonomy in their critical functions, specifically in the selection of a target and the subsequent attack of the chosen target (2016: 1). HRW agree that a fully autonomous weapon would necessarily have the ability to identify targets and trigger an attack by itself (O'Connell, 2014: 226). As can be gleaned from the above definitions, an AWS can fulfil two functions: it can select a target by itself and it can engage that chosen target of its own accord, without human involvement at any stage. The ability to decide for itself is what distinguishes an AWS from other target-triggered weapons systems [51] and what constitutes its designation as autonomous.

[49] Failure can include the weapon system ceasing to function, functioning erratically, or in a manner that is completely outside of its normal scope of functionality.
[50] These levels will be clarified in the section 3.4 "Reframing Autonomous Weapons Systems"
[51] Examples of target-triggered weapons include landmines and booby traps.

The autonomy that makes the weapon what it is consequently removes the judgement of a human agent from the decision to apply lethal force. This separation is the source of both the advantages and disadvantages that AWS offer, which I will discuss more in the subsequent sections. Despite automation of weapons being a controversial issue, there is still a drive towards weapons automation, where the potential benefits of AWS are the force behind this drive. Regardless of the potential benefits, or harms that this technology could offer, it seems inevitable that it will exist. Asaro (2016: 11) believes that rather than framing the discussion regarding AWS as a "utility calculation" where we exhaustively argue about the prospective advantages versus the possible disadvantages, we should recognise that the adoption of AWS in armed conflict is imminent, and thus we should spend our time more productively by discussing the areas where AWS fall short and by deliberating on matters of regulation and arms control. Asaro's sentiment is that we are on a certain path, and it is best to prepare for what is coming rather than debate what could be.

In order to understand the complexity of the issues surrounding AWS better, I will first discuss the potential benefits of the use of this technology, after which I will consider the most challenging obstacles that need to be addressed in the wide-spread adoption of lethal AWS. It is important to keep the original goals of the development of AWS and its associated benefits in mind when designing arms control measures, so the goals and benefits are not abolished by the regulation.

3.2 Potential Benefits of Weapons Autonomy

There are numerous reasons to adopt AWS as offered by its advocates, which is perhaps why the military is investing so heavily in developing this technology. The US military has noted that the use of AWS would likely reduce casualties in war. AWS will also respond faster and with more precision than human agents. These claims are not unreasonable, as will be discussed below (Arkin, 2013: 1). As mentioned in the introduction, automating weapons systems reduces the manpower a military operation requires, and thereby fewer human soldiers are required to go into active combat zones. This makes soldiers safer and would likely reduce casualties. Furthermore, considering that many robots are faster, stronger, and in some cases,

smarter than humans (*ibid*. 3), they are able to react more rapidly. Time can be critical in when responding to threats in armed conflict. Robots will also presumably be engineered to be more physically capable than human soldiers, and their increased precision would likely reduce collateral damage.

If AWS are indeed capable of reducing casualties, and responding more rapidly and accurately, their use would likely save many lives. The preservation of human life is an explicit goal of IHL and is something that the ethics of war strives towards. Thus if AWS can reliably ensure this in a manner that limits[52] the potential harms[53] they can cause, this technology is worth developing. If one considers all the potential benefits of this technology (which I will shortly do more thoroughly), the only conclusion we can draw is that at present, a call for an outright ban on these weapons would be premature. This reinforces the point that arms control regulation is our best option going forward.

Ronald Arkin is a robot ethicist and a vocal advocate for the utilization of AWS for various reasons that follow. Arkin points out that many existing robots are already faster, stronger, and, in some cases, smarter than humans (2013: 3). These traits facilitate rapid decision making, which could mean the difference between life and death, and make AWS more durable and physically capable compared to a human soldier. This implies that AWS are a more ethical [54] weapon than either human alone or human working with passively autonomous weapons. This positive sentiment towards AWS is not restricted to philosophers, even the military has noted that the use of AWS would likely reduce casualties, as they respond faster and with more precision that human agents (*ibid*. 1). This is no doubt part of the reason why militaries and countries are investing so heavily in developing these weapons. Arkin believes that integrating AWS into war will diminish atrocities and almost certainly reduce

[52] Effective arms control regulation would limit the potential harms associated with the use of AWS.
[53] Briefly, the two most pertinent potential harms of the use of AWS is that AWS will be unable to discriminate between combatants and civilians, and that in such cases there is a lack of accountability.
[54] I am using this term loosely here. The argument can be made that killing is never ethical, but for my purposes I presume that a method of war that results in saving more lives or minimises collateral damage compared to another method, as AWS will arguable do, is more ethical.

civilian casualties and property damage (*ibid.*). I will now examine Arkin's argument more closely.

Arkin first claims that since AWS would presumably be more durable than flesh-and-blood bodies, and since they would likely be engineered in a specific manner, they would lack the drive to self-preservation that humans feel acutely when in peril (2013: 4). Arkin postulates that this would give them the option to react to potential threats more conservatively than human soldiers, which would imply that they would not resort to lethal force out of the propensity for self-preservation. This will arguably translate into a reduction in collateral damage. This same logic is also applicable in situations where it is not possible to establish the identity of the target with overwhelming certainty; unlike a human soldier, AWS will not feel as strong a compulsion to protect themselves (*ibid.*). This means that if there were uncertainty as to whether the target was a civilian or combatant, an AWS with a limited drive to self-preservation would not be tempted to err on the side of caution and engage regardless of uncertainty. Arkin suggests that the abovementioned scenarios could be realised by endowing AWS with a "first-do-no-harm" directive (*ibid.*). This would ensure that AWS could truly assume risk on behalf of non-combatants, a policy that soldiers are supposed to employ but which Arkin believes they do not achieve in practice (*ibid.*).

Arkin also asserts that AWS will eventually be outfitted with of a comprehensive array of mechanical instruments that are superior in the battlefield observation to those of their human counterparts, and the ongoing technological developments in various fields[55], means they will be better equipped to penetrate the "fog of war[56]" than humans ever could be (*ibid.*). The uncertainty humans experience in the "fog of war" lead to them making decisions with imperfect and sometimes mistaken information. The superior ability of AWS to sense their environment, coupled with their lack of psychological weaknesses, implies that the "fog of war" will not affect their decision making. Arkin further argues that AWS are capable of assimilating

[55] These include technological developments in the field of robotic optics, acoustics, radar and seismic sensors.
[56] A colloquial term to refer to the uncertainty in situational awareness and confusion experienced by soldiers in the haze of combat.

more data from a vast number of databases[57] as well as environmental inputs[58] much more rapidly than any human is able to, before engaging in lethal action (*ibid.*). This implies that they will be able to make more informed and more consistent decisions, which translates into fewer violations of IHL and less collateral damage. This plausible speculation is why militaries are so invested in the development of AWS, as well as why there are good moral reasons for considering their controlled implementation.

Arkin further argues that unmanned robotics systems would be engineered in such a way that they would not possess passions like anger or frustration which undermine human rationality in high-stress scenarios (*ibid.*). As a result, AWS would not act out of a desire for vengeance or "lose their cool" and react on the basis of emotion rather than reason. He also believes that the use of AWS will avoid the psychological phenomenon of "scenario fulfilment" that often occurs in stressful environments (*ibid.*). "Scenario fulfilment" is a psychological occurrence that causes people to ignore incoming information that contradicts their previous beliefs, or worse, to misrepresent any new information in a way that furthers their pre-existing goals (*ibid.*). A well-known example of this confirmation bias in military settings occurred in 1988 when the crew of the USS Vincennes, a guided missile cruiser, ignored the beliefs they formed about the target that the cruiser's on-board Aegis[59] combat system identified as being a military one, and approved weapons engagement of an Iranian passenger jet (Harris, 2012: 10). The crew, based on their observations of the targets speed and size, identified the target as a passenger jet, but the Aegis program said otherwise and the crew deferred to the programs judgement and initiated an attack (*ibid.*). In fact, there was so much faith in Aegis' targeting system that the USS Vincennes was the only ship in the water where the crew was allowed to fire on their volition without consulting senior officers (*ibid.*). It is a dangerous prejudice, that a person's faith in AI is so strong that they would doubt their own judgement. This was unfortunately not an isolated incident; similarly in 1996 the Japanese destroyer Yugiri shot down an US Navy warplane, after the Phalanx[60] system misidentified the plane as a target during a training exercise. Luckily, in this instance, the two crew members inside the plane

[57] Such as various military networks and the global information grid.
[58] Supplied by their aforementioned superior robotic sensors.
[59] The Aegis combat system is a naval weapons system that identifies and tracks targets and guide weapons.
[60] A naval based close-in weapons system that defends against anti-ship enemy missiles.

parachuted to safety. The point is that in these instances the human inclination towards scenario fulfilment led towards disaster. AWS, being carefully engineered and resistant to emotional responses in high stress scenarios would not have the same inclination.[61]

Arkin also believes that the use of AWS in tandem with human soldiers has the potential to improve human behaviour, since the AWS have the ability to objectively monitor the ethical behaviour of all parties in the battlefield (Arkin, 2013: 5). At the most basic level, this could be achieved by recording military operations and making these recordings available to review by an ethical board. This seems a reasonable assertion to make, when one considers that in the UK the monitoring of police officers by means of body cameras led to a 93% decrease in complaints against police (Gayle, 2016). The monitoring of police officers with body cameras is comparable to AWS supervising human soldiers, as they are both based on the same idea that people behave better when they are under surveillance. While a similar effect might be achieved by requiring soldiers to wear body cameras, such a requirement would be open to abuse; soldiers could disable their cameras when it suits them. Having AWS in charge of the monitoring ensures that the soldiers do not have the ability to manipulate the supervising process. In a more complex capacity, the AWS might be given the authority to intervene when it recognises a violation of IHL.

[61] In both of these cases, while the AI was mistaken, the ultimate decision to initiate an attack lay with the humans operating the system. Therefore, these two example do *not* serve to strengthen the criticism of that autonomous weapons are inherently indiscriminate. Rather, they reiterate the need to have "meaningful human control" in the development stages as well as in the decision to deploy, and the operation of the weapon throughout its deployment. The issue is that AI can be mistaken, and humans tend to defer to them. This shows the importance of "meaningful human control" and the need to have a competent human supervisor that will exercise control over the AWS when necessary. It also highlights the causal link between the human supervisor and the AWS; if the supervisor fails to intervene when they ought to, they are responsible. The point is that having both an AWS and a human supervisor working in tandem is functionally equivalent to a separation of powers in democracies; each serves to check and balance the other. These relationships will be fully discussed in chapter 5.

Arkin makes another interesting point, the full exploration of which is better left to the field of psychology, but is nonetheless worth noting. He claims that human soldiers are placed into conditions and situations under which no human is designed to function (Arkin, 2013: 2). He seems right, especially when one considers the psychological damage suffered by many soldiers when they return from war.[62] In this context, expecting humans to diligently observe the ROE seems, according to Arkin, unreasonable and improbable (*ibid.*). Thus, he believes that machines, lacking human psychology, would be able to keep to the ROE more consistently than a human soldier would, since their application requires objective rather than emotional assessments. Arkin is right in this regard. Arguably, the act of killing another human being has lasting effects on the killer's mental wellbeing. As a result, the training that soldiers go through is meant to make killing an automatic response to a threat, and it stands to reason that such an ingrained reflex is never wholly unlearned. This puts potentially mentally unstable individuals into the civilian public, which undermines public safety. The use of AWS would no doubt diminish the psychological damage suffered by human members of the military, and thereby will minimise the threat these soldiers pose to the public when they return from combat. Part of the reason for developing AWS is to make humans "safer". This goal is also served not only by minimizing physical danger, but also psychological danger.

Broadly, the use of AWS will potentially result in a more ethical method of waging war, as measured by fewer casualties and less collateral damage, and will keep the general public safer, for all the reasons discussed above. For these reasons, this technology is worth pursuing. Despite all the benefits that AWS could offer, it is still important to proceed carefully and slowly when incorporating this technology into military operations, as there are limitations to how AWS can be safely and ethically used. If there is not a careful consideration of the potential shortcomings of AWS, their limitations will not be accounted for and there will not be any meaningful arms control implemented over them. It is important to forge meaningful arms control measures, as it is highly probable that AWS will be deployed in combat in the future. Therefore it is important to understand the criticisms often levelled against AWS, as this understanding that will make arms control more meaningful.

[62] The most common example of such trauma is Post Traumatic Stress Disorder, which is widespread in soldiers (US Department of Veteran's Affairs, 2015).

3.3 Objections to AWS

There are numerous objections given by the critics of AWS that are often cited in the research on the topic, and among these are four principle kinds that Anderson and Waxman (2013) have summarised neatly. Firstly, critics claim it is impossible for an AWS to adhere to the principles of International Humanitarian Law. Secondly, there are no suitable candidates for attributing responsibility to when an AWS commits a war crime. Thirdly, it is amoral to remove humans from the decision to apply lethal force in a combat situation. Lastly, the use of AWS unacceptably lowers the "barriers to war" and make the resort to force more likely. The first two objections warrant further discussion, so the subsequent chapters will be devoted to them. The other two I find less convincing, so I will only briefly discuss them in this section.

The first objection to AWS is that it is impossible to programme AI to exhibit the characteristics required to satisfy the standards for the use of permissible lethal force as set out in International Humanitarian Law (Anderson and Waxman, 2013: 14).[63] Anderson and Waxman note that those in favour of an outright ban on AWS raise this issue the most (*ibid.*). These critics believe that no machine can exhibit the compassion, empathy, and sympathy that they claim make humans irreplaceable in the lawful decision to apply lethal force. Thus the first objection that I will deliberate on (in the subsequent chapter) is the guiding principles on the use of permissible lethal force in combat situations. At this juncture, it is worth noting that this position is hard to hold, since the research on machine intelligence, though currently seemingly in favour of this objection, is only a few decades old, and as such there is no basis on which to make such a sweeping generalisation (*ibid*. 15). The rebuttal of this generalisation is even stronger when one stops to consider that a similar argument was made to claim that a computer would never be able to perform the strategy-forming patterns necessary to beat a human chess champion (O'Connell, 2014: 232), which has been disproven in numerous

[63] IHL is established in several international agreements, including the Geneva and Hague Conventions, and part of it is the regulation of conduct in armed conflict (International Committee of the Red Cross, 2004). All signatories to the various treaties that comprise IHL – which are essentially all the countries that are members of the United Nations – are bound by it.

instances.[64] Furthermore, when one stops to consider the place of human emotions on the battlefield, it is not clear that they are always desirable; passions like fear, anxiety, anger and the like, often intensify the psychological shortcomings of human soldiers that can cause them to violate the ROE (Anderson and Waxman, 2013: 15). Nevertheless, just like any other weapon, AWS will have to be able to be utilised in a manner that adheres to International Humanitarian Law and the principles of proportionality and distinction, part of which involves exhibiting certain characteristics deemed intrinsically human (but I will expand on this in the following chapter).[65]

The second major criticism is related to the previous objection pertains to the legality of the use of AWS. The objection is that even if an AWS could comply with International Humanitarian Law requirements regarding permissible lethal force, there will be some instances where they do not perform as intended and in those scenarios, it will be difficult to assign blame for any violations of IHL that may occur. Critics who favour this position assert that a weapon that is autonomous (more specifically, a weapon that necessarily removes human authorisation from the decision to utilise lethal force in combat) is objectionable because it undermines the possibility of holding anyone accountable for any actions that may violate IHL (Anderson and Waxman, 2013: 16).[66] Anderson and Waxman hold that this objection is raised by those who believe that according to the laws of war, responsibility for crimes is established in a manner similar to that of establishing culpability in criminal law, namely on the basis of mental intention and direct causation (*ibid*. 17).[67] This objection seems like a significant obstacle to the use of AWS, but as will be shown in Chapter 5, it is not as insurmountable as it seems. Briefly, in some instances it will be possible to pin point an individual human decision maker (or group), that acted *mala fide* or was grossly negligent either in the design of the weapon or in the decision to deploy it, and it will be possible to hold that individual (or group) responsible, much like in normal instances of criminal law (*ibid.*). However, as I will discuss

[64] The most famous case is Deep Blue's 1997 victory over Gary Kasparov, the implications of which I consider in Chapter 5.
[65] Intrinsically human characteristics include empathy and compassion.
[66] Actions that violate IHL requirements are generally considered to constitute a war crime.
[67] Mental intention refers to the subject's internal mental states. This was briefly discussed in the previous chapter. Direct causation requires a direct causal link between the action and the criminal consequence. This matter will be dealt with further in the Chapter 5.

in Chapter 5, according to the laws of state responsibility[68], accountability usually lies with the state or the party to the armed conflict and not with a single soldier or commander (*ibid.*). Presumably this will also be the case with regards to states that decide to authorise the use of AWS. Furthermore, Anderson and Waxman point out that establishing individual criminal liability has never been an explicit mechanism of accountability in the military, and it would be inconsistent to hold to this bias in the case of AWS while not doing so in other instances (*ibid.*), and I am inclined to agree with them, since the alternative is logically inconsistent. Currently, attributing responsibility is governed internally by the organisations that would use AWS; the US DoD directive emphasises the need to train human operators to utilise this weapon in a manner compliant with international law, in order to meet the requirement of "meaningful human control", and to be able to recognise if the situation is one that AWS is suited to (*ibid.*). Realistically, the human will simply decide to activate the system, where from that point onwards the system will be permanently activated during that encounter. Human decision makers will have to make the decision to deploy the AWS in every new encounter, and those decisions will have to be justifiable under the laws of war and the ROE (*ibid.*). This would mean that at least some degree of accountability rests with the human operator who elected to deploy the weapon, or refrained from intervening despite recognising a malfunction. Obviously, leaving matters of responsibility to be solely determined internally by the military is not desirable, as there is no guarantee that guilty parties will be held accountable. In the chapter on responsibility, I will examine the above-mentioned issues more closely.

The third core objection against the use of AWS is based on the belief that it is immoral to remove humans from the decision to employ lethal force in combat situations. These critics believe that no machine could ever hope to replicate the judgement exhibited by a human moral agent and the associated understanding of the significance of the act of killing; therefore leaving this decision up to a machine would be immoral and we should avoid doing so at all costs (*ibid.* 15). It is difficult to debate this point since it is a moral belief that you either accept or not. However, I believe that it is possible to refute the call to ban AWS based on this principle if one examines the campaign to ban antipersonnel landmines. The reason for the ban on landmines was not based on the fact that there was no human agent pulling the trigger; rather,

[68] The laws of state responsibility govern instances when a state or, in particular circumstances, state representatives violate an international agreement.

it was because the weapon fails to satisfy the principle of distinction[69], meaning there is no meaningful decision made on target selection. This objection is therefore merely a reformulation of the first objection. Thus, by implication, if the weapon could sufficiently discriminate between combatants and civilians, there is no reason to ban AWS based on the sentiment that a human ought to be pulling the trigger. Regardless of whether one accepts this principle, we are already turning more life or death decisions over to machines, from driverless cars to auto-pilot on planes and automatic robot surgeries, because these tools are more convenient and generally safer (*ibid*. 16). The fact that we are already turning these decisions over to machines highlights the inevitability of the use of AWS and shows that this objection is not enough to slow their development. Based on the above discussion, I do not believe that this criticism is strong enough to warrant an in-depth examination, since it ought to be sufficiently dealt with in the discussion on whether AWS can satisfy IHL.

The fourth and final main objection to AWS is that minimising risk to human soldiers lowers the "barriers to war" and the disincentive to engage in armed conflict is diminished (*ibid*. 18). In other words, these weapons diminish the mortal danger faced by soldiers, and in doing so they reduce the "cost" of war, effectively disrupting the usual means of conflict (*ibid.*). This is a strange objection, yet it comes up a lot amongst the critics. It is peculiar because it implies that any action that minimises risk to soldiers by protecting them, essentially lowering the loss of human life, is undesirable. If you accept this claim, then implicitly we ought to make war as costly in terms of human life and collateral damage, and as dangerous to soldiers as possible. This is a ludicrous position to hold, so this objection must be rejected. Moreover, one of the guidelines of the laws of war, which will be discussed in the next chapter, is that in a just war[70] parties to war ought to employ the most conservative approaches and means of war available. Therefore, it is actually an explicit goal of International Humanitarian Law and the laws of war to reduce casualties, thus one could maintain that we are legally required to adopt AWS, since they arguably further this goal.[71]

[69] Briefly, the principle of distinction entails differentiating between persons who are directly participating in hostilities and civilians that are entitled to protection. This principle is discussed in more detail in the next chapter.

[70] A war that is considered legal and is sanctioned by the international community.

[71] The case that the use of AWS reduced causalities was discussed in the section on the potential benefits of AWS.

In an attempt to assuage fears such as the ones discussed above, the US DoD Directive 3000.09[72] states that only human-supervised AWS may be used to select targets (Cummings, 2014: 2), and that these systems will be engineered in a way that maintains "meaningful human control" over them, or that facilitates the application of what is termed "appropriate levels of human judgement[73]", in all operations (Sharkey, 2012: 2). Indeed, keeping a human involved at some stage of the decision-making process with regard to employing lethal force is a broad strategy that militaries have used to try overcome the criticisms levelled against AWS. Problematically, this strategy only works if the human involved has clearly defined and meaningful control. However, generally there is no further explanation of the mechanisms relating to this strategy – beyond simply stating that there will be a level of meaningful control. This vagueness needs to be clarified if implementing meaningful human control is going to be a legitimate strategy for AWS regulation. Furthermore, if "meaningful control" is to be used as a method of arms control, it needs to be formalised in an international setting, as opposed to it being a restriction that militaries impose on themselves. This will ensure that the military cannot revoke the principle when it suits their purposes. Consequentially, I will now turn to a discussion by Peter Asaro, where he gives an account of what could constitute "meaningful human control".

3.4 Meaningful Human Control

Asaro begins by examining the origin of the standard of "meaningful human control". He writes that the idea of "meaningful human control" seems to be an emerging principle in the debate over AWS, as it keeps humans "in the loop" of the lethal decision-making process (Asaro, 2016: 12). As mentioned, keeping humans "in the loop" is one of the ways the opponents of AWS suggest we comply with International Humanitarian Law. Asaro believes that the argument could be made that the principle of "meaningful human control" has always existed in the law, but that it has never been expressly articulated (*ibid.*). Thus, the idea of "meaningful human control" has never been a legal issue before. He bases this assertion on the point that there is a broad-based acceptance, both in the military and in the institutions of

[72] This was discussed in an earlier section in this chapter.
[73] This phrase is synonymous with the concept of "meaningful control".

international law, that any instance of permissible killing needs at all times to occur after "appropriate levels of human judgement" have been exercised and on the basis of "meaningful human control" over the weapons involved (*ibid.*). In order to prove his claim, he looks at how the introduction of weapons like land mines and cruise missiles was met with unease at what they represented for the future of warfare, and claims that this unease was due to the implicit belief that people ought to be directly involved in the kill chain (*ibid.*). Asaro seems justified in his assertion; people have traditionally been necessary in the activation and use of arms, and weapons like booby-traps and antipersonnel landmines that challenged this norm where met with protest because of the lack of human decision making involved in the application of lethal force.[74]

Today there are strict restrictions on the use of weapons that are insufficiently discriminate[75], but they do not yet apply to AWS. Since the formalization of arms control measures is a lengthy process, as mentioned in the introduction with regards to the battle against cruise missile proliferation, it is necessary to start this process as soon as possible, and this includes defining "meaningful human control". Military ethicists have to address moral considerations regarding new weapons, and then these deliberations need to be brought to policy makers, who in turn codify regulations and formalise legislation, which when passed, finally culminates in arms control (*ibid.*). Regarding AWS, this process is still in its early stages. As stated above, Asaro holds that the principle of "meaningful control" has always existed in the laws of war, albeit implicitly, since previous generations of weapons required a human operator (*ibid.*). We have always assumed that this norm existed but this principle has never been formalized in legislation (*ibid.*). Asaro believes the reason that this idea is being expressly articulated today, is because AWS pose a greater challenge to the assumptions that we previously held, and just like some weapons in the past [76] they have prompted moral

[74] As discussed in the subsection on criticisms of AWS, the issue here is not the specifically the lack of a human agent pulling the trigger, so to speak, but rather the lack of the application of the principle of discrimination, that is the issue.

[75] Booby traps and antipersonnel landmines are both clear examples of weapons that are considered inherently indiscriminate.

[76] Weapons that challenged the norms of warfare, from weapons like the long bow to nuclear ordnances.

considerations based on the principles of humanity and the dictates of public conscience (*ibid.* 13).[77]

Asaro believes that is important to comprehensively articulate the concept of meaningful human control, otherwise AWS may end up without any meaningful control, where the risk of unintended consequences is high and where responsibility for the decisions to use lethal force is unclear. Such clarification would help eliminate any ambiguities and would prevent humans from feigning ignorance in order to escape responsibility (2016: 12). Under such circumstances it would be easy for states and individuals to misuse AWS and escape accountability. Asaro argues that "meaningful control" has two elements to it that need to be examined: firstly, the concept of control, and secondly the qualifier "meaningful" (*ibid.*). Both of these need to be clarified in order to ensure effective human supervision over the application of lethal force by AWS.

Firstly, Asaro states that we need to clarify what constitutes control (*ibid.*). He writes that at minimum, "control" indicates that the results delivered by the weapons system ought to be reasonably consistent and predictable such that any human operating the system is able to govern it (*ibid.* 14). Furthermore, the outcomes produced by the system ought to be in line with the objectives envisioned by the operator (*ibid.* 12). When such results are not similar enough to these goals, it is reasonable to conclude that the operator is no longer in control of the system (*ibid.*). His definition is consistent with the normal conceptualisation of "control", namely to dominate or direct something. I will take "control" in this context at minimum to mean a human is able to channel or exercise authority over the system if she/he chooses to do so. In other words, a system that is under some form of control must have the protocols in place that allow it to submit to a human that is recognised to be in a position of authority. The recognition of the authority of the human operator leads to the qualifier that needs to be expanded upon, namely "meaningful".

[77] These two concepts are a part of the Marten's Clause, which always comes up in disarmament debates, and it is elaborated on in Chapter 4.

Asaro seeks to clarify is what it is that qualifies the control that a human operator has over an AWS as being "meaningful" (*ibid.*). He believes such clarification is crucial in order to prevent deployment of AWS where the operator is reduced to merely being used as an apparatus to sanction the application of lethal force (*ibid.* 14). Asaro gives an example to illustrate his concerns: if the human operator has been ordered to approve lethal engagement every time a prompt appears (possibly provided by the AWS), he is technically in control, since the AWS technically awaits the operator's approval, but it is doubtful that such control is meaningful (*ibid.*). In this example, the human operator is no better than an automated system, and there is no meaningful and considered human judgement involved in the application of lethal force. Furthermore, Asaro believes that a part of such control being "meaningful" is that the operator accepts responsibility for the decision to deploy the weapon and for any decision to abstain from intervening in its operations (*ibid.*). I would tentatively agree with Asaro here; if the human has orders to approve all targets selected by the AWS, they are essentially automated systems themselves. Asaro argues that one can only be held responsible for a weapon that one has meaningful control over, as the two concepts are interconnected. As such, Asaro's position is that any human that operates merely as an approval mechanism has diminished responsibly. I would agree with him, to a certain degree, on this point; the human operator has less responsibility when they are under orders to defer to the AI's judgement. However, I would also point out that this is exactly the type of logic that facilitates the commission of war crimes by large organisations. I will return to this argument in Chapter 5, for now it is sufficient to note that "meaningful control" and responsibility are intertwined, and if the human operator merely approves all AWS decisions, there is neither meaningful control nor full moral responsibility.

Asaro takes his argument further; he argues that "meaningful control" is also legally required in order to satisfy the IHL requirement that the lawful application of lethal force be intentional. In other words, any instance of permissible lethal force must be employed knowingly and for a reason (*ibid.*). He is partially correct in this assertion, since the lawful application of lethal force does require intentional and purposeful action; I will elaborate on the conditions for the permissible use of lethal force in the next chapter.[78] However, I believe

[78] Chapter 4 discusses International Humanitarian Law.

Asaro understands "intentional" in a more philosophical sense[79], as opposed to the manner it is normally used in IHL. Asaro's understanding of intentionality requires that the reason behind an action or the significance of the act be understood by the perpetrator of the act (*ibid.*). I believe this understanding is based on associating intentionality with mental states, or "aboutness", as discussed in the previous chapter. However, the presumption of understanding the significance of the act does not exist in law. What is required is that the decision maker *made* the decision, and that the application of lethal force was on purpose and not accidental.[80] Asaro asserts that while it is possible for AWS to be designed to respond to lethal situations according to varying governing conditions[81], it is debatable whether they are capable of comprehending the significance of their actions. It is for this reason that he and many critics argue that AWS are intrinsically incapable of legally making the decision to kill (*ibid*). [82] Essentially, his point is that a machine lacking intentionality would not be able to lawfully apply lethal force. I believe that given the legal use of the term, this is not true. All that would be required is that the AWS *made* the decision to engage, and did so purposefully. Furthermore, as I will discuss in the subsequent chapter, if the machine is under the kind of "meaningful control" discussed above, the argument could be made that the human who deployed the AWS, or failed to intervene, is the intentional and responsible agent relevant.

Nevertheless, what Asaro is attempting to establish is that any legal use of AWS must be done under circumstances of meaningful control, not only to satisfy this new emerging principle, but also because according to IHL any instance of permissible lethal force must be intentional and meaningful. I agree with his argument. In order to meet the International Humanitarian Law requirement of intentionality there must necessarily be a human in meaningful control of an AWS. Asaro also believes that reaching consensus on what constitutes "meaningful human control" would be a source of direction on how systems should be engineered in order to keep humans in- or at least on- the loop when it comes to lethal decisions (*ibid*. 15).[83] The ICRC agrees with this notion; they assert that for an AWS to be ethically and

[79] The philosophical understanding of "intentionality" was discussed in chapter 2. Briefly, it refers to an internal, subjective quality of consciousness.
[80] The exact requirements for the lawful application of lethal force will be clarified in the next chapter.
[81] In other words, they can act and react the way a human would.
[82] This is not the only reason that critics think so, and additional reasons will be discussed shortly.
[83] A human who is on the loop is not involved in the decision to kill, but merely passively supervises an AWS and interferes when necessary.

legally acceptable, the AWS ought to be under human control during development and programming, during their deployment, and throughout the operation of the weapons (2016: 3). It is difficult to imagine how development and deployment could be guided towards the goal of maintaining "meaningful human control" without consensus on what meaningful control involves, thus I agree that reaching consensus is critical.

In order for the development of AWS to be morally acceptable, there must be, at the very least, the kind of meaningful human control Asaro proposes over the development and use of AWS in the first few generations. This control must allow the human operator to observe, guide, and, if need be, overrule the system, and to employ their own judgement such that they are causally responsible if they fail to exercise that judgment. Whether or not such control can effectively be given up entirely at a later stage is something that can only be determined after AWS under human control have been extensively supervised over a significant period and AWS have been endowed with full personhood.[84]

3.5 Summary

When examining AWS there is an apparent ambiguity in the employment of the terminology and, more specifically, confusion between the terms "automated" and "autonomous". "AWS" is a broad term used to denote a range of weapons system that operate without human supervision. While it is necessary to maintain a broad definition to accommodate a wide scope of AWS, this understanding does not accurately define the concept since it includes automated systems. Thus the concept needs further refinement. When examining the various definitions offered by the United States Department of Defence, and other concerned groups like HRW and the ICRC, it becomes apparent that AWS are defined by two critical functions: firstly, by the ability to independently identify and select a target, and, secondly, by the ability to engage the target selected in the first function. The ability to perform these two functions without human interaction is what differentiates an autonomous system from an automated one.

[84] The reason for this is that if the AWS is legally considered to be a person, they are legally interchangeable with other human agents.

As I will discuss in the following chapter, in order to be legitimate, the two functions that characterise AWS, which constitute important stages of the decision to employ lethal force, must occur in compliance with International Humanitarian Law, specifically with the principle of discrimination and the principle of proportionality. The current method employed to keep the use of AWS compliant with International Humanitarian Law is to restrict their autonomy by means of "meaningful human control". This concept is also one that needed further elaboration. We saw that Asaro argues that control is constituted by the system delivering the results intended by the operator, and I further pointed out that this includes the operator's ability to guide, direct and overrule the decisions of the AWS if need be. According to Asaro, control is meaningful when the operator is able to exercise his or her own authority over an AWS (and not merely functioning as an approval mechanism for the AWS' actions) such that he or she could be held meaningfully responsible if he did not exercise his or her own discretion.

Asaro further believes that meaningful human control entails that any instance of lawful killing using an AWS should be intentional; however, I argued that he mistaken in his understanding of intentionality. Nonetheless, his overall position that meaningful human control is something that we should endeavour to realise is valid and one that I accept. Maintaining that meaningful human control over AWS should be a standard legal requirement is valid and something we should endeavour to realise. Furthermore, if meaningful human control is to be effective, it should be considered in the engineering and development phases. In other words, AWS ought to be designed in a manner conducive to meaningful human control.

I have discussed some of the potential benefits of AWS, including Arkin's postulation that given the superior detection technology and the lack of human passions like anger and fear, AWS could perform more ethically and conservatively than human agents could. I therefore concluded that a call for an outright ban is premature. But the adoption of AWS faces a number of obstacles that need to be addressed before their widespread use. As I pointed out above, opponents of these weapons often cite four types of criticisms to these weapons, some of which are easily dismissed and others not. The criticisms that are not easily dismissed are what I will

focus on addressing in the chapters to follow. The first issue to address is when it is permissible to use lethal force and how AWS will recognise these situations, which is the subject of the next chapter. The second concern relates to matters of responsibility for the decision to employ lethal force, which will be discussed in Chapter 5.

4. Permissible Lethal Force

As mentioned in the previous chapter, while there are some good reasons for adopting AWS, there are also some important legal and ethical considerations to take into account. The use of these weapons must be strictly regulated. One of the purposes of this chapter is to examine the instances when lethal force is permissible according to International Humanitarian Law in order to gauge when the deployment of AWS might be appropriate, and to determine what requirements the use of AWS would have to adhere to. Thus the bulk of the discussion will examine the existing laws of war to determine which, if any, of the existing laws are sufficient to govern the use of AWS, where the law needs to be amended, and when entirely new legislation is necessary.

One of the chief criticisms against the use of AWS is that they are inherently incapable of satisfying the requirements for the lawful application of lethal force as set out in International Humanitarian Law. Any instance of lethal force that does not conform to these requirements is deemed unlawful and thus criminal. In some cases, the unlawful application of lethal force can constitute a war crime. Any instance of unlawful killing generally necessitates that the perpetrator be held responsible, thus examining the lawful application of force is a necessary precursor to any discussion on responsibility. In order to determine whether the assertion that AWS are intrinsically incapable of conforming to the requirements for the lawful use of lethal force has any merit, it is necessary to examine the circumstances that permit the use of lethal force. Arguably, AWS would need to satisfy the same criteria as human soldiers do, as holding them to a higher standard seems unfair. Thus the focus of this chapter will be on the legal principles that guide the use of lethal force in international law.

While AWS could potentially be deployed under a range of contexts under the general heading of law enforcement and peacekeeping, or even in the private sector, the majority of the debate has focused on the use of AWS in armed conflict (Human Rights Watch, 2014: 7). Thus, I will focus on the instances of the permissible use of lethal force in armed conflict.

Though I will only be explicitly discussing international law as it pertains to armed conflict, this does not mean that the outcome of these discussions are limited to armed conflict; the norms established in international law tend to guide domestic legislative standards, and as such, any conclusions reached regarding the use of AWS in armed conflict should be translatable to their lawful use in domestic and peacekeeping settings. This translation of international law into domestic policy is facilitated by the fact that International Humanitarian Law (IHL) applies in all circumstances, from armed conflict to peacetime, within both domestic and international settings (O'Connell, 2014: 230). The source of the regulations that guide the lawful use of lethal force in armed conflict is the various bodies of IHL.

The purpose of this chapter is not only to establish what criteria needs to be met for an application of lethal force to be lawful, but also to determine whether the existing legal system is able to adequately regulate AWS. As mentioned in both the introduction and the previous chapter, forging arms control takes time, and if significant changes to existing conflict law needs to be made, it is something that we should seek to institute sooner rather than later. In general, arms control regulation is codified in treaties and multinational agreements, and so another purpose of this discourse is to guide nations and offer suggestions for definitions and practices that could easily yet effectively be employed in order to forge strong arms control regulations over the use of AWS. The point of departure for investigating the applicability and adequacy of the existing legal system is the examination of the relevant bodies of legislation that form international law.

The Rome Statutes of the International Court of Justice[85] recognizes four categories that constitute international law: conventions, customs, general principles and judicial decisions (Asaro, 2016: 10). More specifically, the permissible use of lethal force in armed conflict is governed by the subfields of IHL that form the Laws of War, the Hague Conventions, the Geneva Conventions of 1949 and their Additional Protocols of 1977, international

[85] The Rome Statue is a part of the Charter of the United Nations. This treaty's primary function was to establish the International Criminal Court. While the International Court of Justice is the official judicial branch of the United Nations, the International Criminal Court is an intergovernmental tribunal that has jurisdiction over contravention of IHL, or international crimes.

customary law as well as existing jurisprudence (O'Connell, 2014: 230).[86] It is a daunting task to examine all of these, so the focus here will be on the principles, rather than specific documents, that constitute the Laws of War, namely *jus ad bellum*, or the "right to war", and *jus in bello*, or "just war", since these inform the laws that govern the resort to lethal force in armed conflict (*ibid*.). The first outlines the circumstances under which a country is permitted to resort to or initiate armed conflict. The second discusses acceptable conduct in the exercise of force during armed conflict (also called the Rules of Engagement), part of which requires conforming to the principles of proportionality and distinction (*ibid*. 229). It is important to note that just conduct in armed conflict is a separate issue from the just initiation of war, and even if the initiation of hostilities does not conform to *jus ad bellum*[87], conduct during hostilities is still expected to conform to *jus in bello*. In other words, if a country initiates an unjust war, the soldiers are still expected to adhere to the principles of IHL, otherwise they may be charged with the commission of war crimes.

The two most important pillars of IHL, and principles embodied in both *jus ad bellum* and *jus in bello*, are the right to life and the right to human dignity. Article 1 of the Universal Declaration of Human Rights[88] articulates the right to human dignity, which means that every individual has the inherent right to be valued and treated ethically. Article 3 of the same document guarantees every human the right to not be arbitrarily be deprived of their life (Human Rights Watch, 2014: 8). These rights are the bedrock of IHL; they are supreme rights[89] and as such are non-derogable, even in public emergencies (*ibid*.). These rights are embodied in the Laws of War and the Rules of Engagement, and they are important to bear in mind, as any lawful use of AWS cannot contravene them. In other words, the use of AWS must always be against a legitimate target and the application of force must not be excessive. This is equivalent to not arbitrarily depriving someone of their right to life. Currently, these determinations must be made in each instance by human soldiers, and AWS will have to make the same determinations if they are to lawfully employ lethal force.

[86] Theory or philosophy of law.
[87] It should be noted that initiating an armed conflict without conforming to *jus ad bellum* constitutes a war crime.
[88] This document is a part of IHL, and offers protections to people even during armed conflict.
[89] A supreme right is a prerequisite right and all other rights are dependent on it.

Part of what makes the use of a weapon lawful, is that it is able to be used in a manner that is discriminate and proportionate. These requirements are designed to ensure that no one is arbitrarily deprived of their right to life and dignity. Asaro notes that these constraints themselves are not enough to provide a strong ground for banning a weapon, as any weapon could potentially be used in an indiscriminate or disproportionate manner (2016: 7). Only weapons that are considered to be intrinsically incapable of being used in a discriminate or proportionate manner ought to be banned absolutely based on these requirements (*ibid*.). For those weapons where there are some circumstances in which their use could be indiscriminate or disproportionate, their use ought to be restricted so that they can only be used in circumstances where their use is discriminate and proportionate, in compliance with IHL (*ibid*. 7-8). For example, unguided missiles can be assessed as discriminate or indiscriminate depending on the circumstances of their use, and, as such, their use is explicitly regulated and firmly restricted. In other words, only weapons that *de facto* exclude the possibility of discriminate and proportionate use ought to be banned outright,[90] and weapons that simply have the potential to be used unlawfully or unethically should be restricted and subjected to arms control.[91] This attitude is the prevailing consensus amongst academics who have examined the use of AWS.[92] Enforcing arms control would also be consistent with holding AWS to the same standards that other weapons or human soldiers are held to. AWS are not inherently incapable of discriminate and proportionate use, and as I examine the principles guiding the lawful use of lethal force, it will become apparent that AWS can be used in a manner that complies with IHL. Before I review all of the conditions that have to be satisfied in order for the application of lethal force to be lawful, it is necessary to discuss the Martens Clause, which is an aspect of IHL which is almost always cited in disarmament contexts. This discussion serves to reiterate that there are no legal grounds for a ban on the use of AWS and so reinforcing the claim that regulation is the best, and indeed, only way forward.

[90] The only examples of such weapons that I can think of are weapons of mass destruction, like the atom bomb.
[91] Most weapons have this potential, therefore most weapons have restrictions on their legal use.
[92] Academics that hold this position include Peter Asaro, Anderson and Waxman, whose views are discussed throughout this dissertation.

4.1. The Martens Clause

The Martens Clause is a part of IHL which is often invoked in disarmament contexts because it embodies the ideals of public conscience and humanity, thus it is relevant in any discussion of weapons prohibition and regulation. It first appeared in the preamble to the Hague Convention II on the Laws and Customs of War on Land in 1899, but the latest formulation was in the 1977 Additional Protocols to the Geneva Conventions (Asaro, 2016: 4). In its most recent version, it reads "in cases not covered by the law of force, the human person remains under the protection of the principles of humanity and the dictates of public conscience" (*ibid*.). The most commonly held interpretation of the clause is that acts are not legal or permissible simply because they are not explicitly prohibited by law. For AWS, this means that the lack of specific regulations against their use does not mean that their use in an unconstrained manner is necessary lawful or permissible. The Martens Clause refers to two ideas that need further examination, the first is the "principles of humanity" and the second is the "dictates of public conscience" (*ibid*.). Neither of these is made explicit in the clause, so Asaro attempts to explicate them further.

Asaro writes that the reference to "principles of humanity" has two possible interpretations. It could potentially be understood to be equivalent to the specific clauses embodied in the Universal Declaration of Human Rights. The other way to interpret it is as a broader ethical ideology that forms the basis of our shared notion of humanity, which is far more open to interpretation (*ibid*. 5). Asaro believes that some have a radical perspective on the goals of IHL, specifically the belief that the explicit goal of IHL in armed conflict is to categorically minimise the threat to civilians. He believes that rather, it would be more realistic to understand the purpose of IHL as an effort to safeguard a broader notion of humanity amidst the inhumanity of war (*ibid*. 9). This seems prudent, as the law allows for a certain, proportional, amount of collateral damage. In other words, it is not unlawful to take an action that brings about the death of civilians, if their deaths are proportionate to the gains brought about by that action. This will be discussed in more detail in the section on proportionality, but this sentiment returns us to the Martens clause and the "principles of humanity" (*ibid*.). This would mean that if AWS could be used in a manner that minimises, but does not necessarily preclude, civilian casualties, their use would be consistent with the "principles of humanity" and the Marten's

Clause. And given Arkin's discussion on the potentially superior abilities of AWS in the previous chapter, the use of AWS in a manner consistent with the Marten's clause is possible. Thus, there is no compelling reason to ban AWS based on the "principles of humanity".

Asaro further states that the "dictates of public conscience" are generally considered to be roughly equivalent to public opinion; however, strictly speaking, it is not the same thing (2016: 5). He argues that since public opinion is subject to manipulation by strategic communication and propaganda that it is improper to equate the two completely; however, they do overlap. He points out that the crucial difference between public opinion and public conscience is that the latter has an explicitly moral component that the former lacks (*ibid*. 6). Opponents to AWS claim that the use of AWS will violate the "dictates of public conscience", and, therefore, we should not consider using them (Arkin, 2013: 7). Asaro examines what "public conscience" is in light of AWS and identifies three anxieties that opponents within the general public have regarding the use of AWS. Firstly, there is concern over the risk to civilians that these weapons pose (Asaro, 2016: 7). Secondly, many are worried about the implications of the use of AWS and the introduction of artificial moral agents on human rights and dignity (*ibid*.). Lastly, as mentioned in the previous chapter, there is unease over how the use of AWS will transform the use of violent force in warfare and "lower the barriers" to war (*ibid*.). I have already dismissed this point. Arkin, the AWS advocate discussed in the previous chapter, concluded that until the Martens Clause and the "dictates of public conscience" are more clearly defined, it cannot legitimately be used as grounds to ban AWS (2013: 7). This is because a ban based on a vague principle opens the floodgates to banning anything based on that same vague principle. The "dictates of public conscience" could be used to declare anything controversial as unethical and ban it without having a compelling reason to do so.

Despite the Marten's Clause often being cited as grounds to ban weapons, neither of its two principal concepts are strong enough to warrant an outright ban on AWS. Thus the best option opponents are left with is arms control regulation. This is also an option favoured by proponents; despite being in favour of AWS most advocates recognise the potential harms inherent in this technology and recognize the need for arms control. If AWS are to be used, their use must be compliant with the international standards for the permissible use of lethal

force. Therefore, it is necessary to move on to the guidelines human soldiers are provided with regarding the permissible use of lethal force, in order to understand when AWS can make these same lethal decisions in compliance with IHL. As mentioned earlier, the resort to armed conflict is governed by *jus ad bellum*, and the use of lethal force in armed conflict is governed by *jus in bello*. These two braches of IHL will govern when the use of AWS is acceptable, and when lethal force can be applied.

4.2. *Jus ad Bellum*

International Humanitarian Law prohibits the resort to lethal force except in specified circumstances. These circumstances are outlined in the branch of conflict law referred to as *jus ad bellum*, or the "right to war". Certain weapons, like cruise missiles, are only permitted to be used in armed conflict situations. Currently, the norm is building that the use of AWS will be similarly restricted, thus it is necessary to recognise which situations will allow for the deployment of AWS. In peacetime, the police and government may use lethal force only to save lives immediately, and no innocent bystanders may be killed (O'Connell, 2014: 229). This is why it is currently possible to use highly autonomous weapons systems[93] in a defensive capacity during peacetime and still comply with international law, since reacting to an attack in self-defence will presumably save lives immediately.

The use of lethal force is also restricted in instances of domestic and international armed conflict; governments may only resort to lethal force when threatened by organised insurgents on their own territory or when attacked by significant force from abroad (*ibid*.). Only a legitimate authority recognised by the international community (for example, a sovereign state) has the right to wage war and a war can only be deemed just if waged by such an authority (Lazar, 2016). A war is also only considered just if the aims of that war are just (*ibid*.). In most cases, "just cause" entails the reestablishment of peace, and excludes the pursuit of economic interests. Furthermore, a just war must have a reasonable chance of being successful (*ibid*.). That means there needs to be a strong basis for concluding that the aims of a just war are achievable. In other words, hostilities are impermissible if it is questionable that such actions

[93] Examples have been mentioned in the previous chapter.

will realise the aims of the war. The initiation of hostilities must also be proportionate[94] to the threat that prompted a hostile response, and armaments like missiles, bombs and the like, which are considered less discriminate, are only permissible for use in armed conflict hostilities, or to save lives immediately (O'Connell, 2014: 230). The United Nations Security Council also has the authority to sanction military activities with the goal of peacekeeping (*ibid*. 229).[95] This presumably means that the United Nations would be able to sanction the use of AWS for the purposes of peacekeeping operations.

When we apply the restrictions discussed above to AWS, we can see that in contexts where a country has a "right to war", governments and militaries may only sanction the use of AWS if they are doing so to save lives immediately or when they face a significant threat domestically or internationally, and when the aims of their use are considered "just" and achievable or when sanctioned by the United Nations. The restriction on less discriminate weapons will be discussed in the next section, but it is clear that if AWS were considered to be inherently less discriminate weapons, like missiles or bombs, then their legitimate use would be restricted to armed conflict only. I do not, however, believe that AWS are inherently indiscriminate; while some AWS will affect a larger area, others will be extremely precise. AWS are comparable to other advanced ballistics weapons in this regard, and, for the most part, the use of AWS will only be appropriate in armed conflict. However, there will also be some AWS that are sufficiently precise and small scale that can be sanctioned for use outside of armed conflict. These determinations will have to be made on an individual basis, and effective regulation should ensure that AWS are used in situations where the use of weapons of this nature is appropriate. Given that AWS are already being used legitimately during peacetime (in a defensive capacity), it is plausible that AWS could be deployed just as legitimately in armed conflicts if the initiation of hostilities meets the guidelines of Just War principles. Even if they do not, the use of AWS could still be lawful as long as their use is governed by *jus in bello*, or "just war". This branch of law provides the guidelines for the permissible use of lethal force, the two most important being the principle of distinction and the principle of proportionality.

[94] The principle of proportionality will be discussed shortly.
[95] The active maintenance and preservation of peace and security.

4.3. *Jus in Bello* – The Principle of Distinction

Jus in bello refers to the branch of IHL that governs conduct in armed conflict and the resort to lethal force in armed conflict. Any lawful application of lethal force, by humans or by AWS, has to adhere to *jus in bello* and to the principles articulated within this branch of IHL. The two principles that guide the permissible application of lethal force are the principle of distinction and the principle of proportionality. The principle of distinction is based on the rule against inherently indiscriminate weapons and has to do with differentiating combatants and civilians, as set out in Article 54 (b) (4) of Protocol 1 of the 1977 Additional Protocols to the Geneva Conventions (Anderson and Waxson, 2013: 10). One of the strongest criticism against AWS is they cannot satisfy the requirements needed for the legitimate use of lethal force. Specifically, the argument is made that AWS are unable to differentiate between combatants and civilians accurately enough to be entrusted to make reliable and lawful decisions regarding the use of lethal force. This criticism is important to address, because as mentioned, any weapon that is incapable of being used in a sufficiently discriminate manner ought to not be fielded.[96] As mentioned in the previous chapter on autonomous systems, while the claim that AWS cannot discriminate sufficiently is largely true now, there are no grounds for concluding that this will always be the case.

One of the goals of IHL, especially in instances of armed conflict, is the protection of ordinary people who are not directly taking part in the hostilities, or who have ceased to directly take part in hostilities, as well as to delineate the legitimate methods of waging war (Asaro, 2012: 697). As such, soldiers are expected to discriminate between people who are protected by IHL and active combatants. All people who are neither members of the armed forces, nor members of a *levée en masse*[97], or who are excluded due to being *hors de combat*[98], are not considered to be legitimate targets and are entitled to all the rights and protections afforded by IHL, including the right to be protected from hostilities. Civilians can only be considered to be valid targets (and therefore, technically, no longer civilians) if they are fulfilling a role in a

[96] The best example of an inherently indiscriminate weapon is antipersonnel landmines.
[97] A mass uprising.
[98] Literally meaning "outside the fight", this term is used to refer to people who are incapable of performing their capacity to wage war. In other words, people who have surrendered or who are unconscious.

"continuous combat function[99]", or more generally, they are deemed to be members of an organised militia (O'Connell, 2014: 231). This criterion is measured by the acts of the persons involved, and any person who performs an act that constitutes "directly participating in the hostilities" loses the protections recognised under IHL. In order to employ lethal force, any AWS would have to determine whether a potential target is "directly participating in the hostilities" before selecting them as a valid target. There are three components of an act that constitute it as being an instance of "direct participation in hostilities" in armed conflict, and each of these three must be established concurrently to determine that a target is appropriate: (1) the threshold of harm, (2) direct causation, and (3) belligerent nexus (Asaro, 2012: 697). These concepts will now be discussed in detail.

In terms of the first component, threshold of harm, it must be determined that the actions of the potential target are expected to unfavourably affect the military's operations or diminish its capacity[100], or in the absence thereof, the act must be expected to result in the death or injury of persons who are protected against direct attack, or in the destruction of civilian property that is similarly protected (*ibid*.). The relevant determination is based on the likely harm that can be reasonably expected to occur as a result of the act. The threshold will, in general, be reached when the probable harm caused by the act is of a military nature, but it does not exclusively refer to military harm and the threshold can still be reached if the harm is directed against non-military targets, like civilians (International Committee of the Red Cross, 2009: 49). In addition, the threshold of harm does not come into play if a civilian refuses to assist the military; their refusal cannot be interpreted as negatively impacting on military operations (*ibid*.).

After establishing that the act meets the required threshold of harm, it is necessary to establish that the act meets the threshold in one causal step, so that the perpetrator of the act is directly linked to its outcome (Asaro, 2012: 698). This is the "direct" part of the action, and this criterion is formulated so as to only include activities that contribute objectively to the

[99] A term employed by the ICRC to denote those who are participants in hostile activities. For example, soldiers and insurgents.
[100] Examples of such acts would include sabotage or disturbing deployments, logistics and communications, electronic interference and transmitting tactical information.

military defeat of an adversary[101] and not to merely war-sustaining activities (International Committee of the Red Cross, 2009: 51).[102] Admittedly, while some war-sustaining activities could qualify as direct participation in hostilities, they may fail to meet the required threshold of harm, and are so excluded. For example, financing hostilities are war-sustaining, but they do not meet the first criterion, thus financers of war are not "directly participating in hostilities". The standard of "direct causation of harm" is important because it stops civilians from being identified as targets based on indirect causation or the mere facilitation of harm, as this would make the scope of legitimate targets unacceptably wide (*ibid*. 52). "Direct causation" usually means "in one causal step", meaning that there must be a clear causal link between the act and the outcome, thus excluding political, financial, and scientific acts from qualifying as instances of direct participation in harm (even though these may all be directly imbedded in the hostilities) (*ibid*. 53). Hence, the complex and collective nature of military operations is accounted for, and this requirement is formulated in such a way so as to ensure that the direct causation of harm is only attributed to those who play an integral role in the act (*ibid*. 54). It excludes those who have causal proximity, in that it does not confuse temporal or geographic proximity with direct causation of harm (*ibid*. 55).

Lastly, any act that in one causal step meets the required threshold still would not constitute "direct participation in hostilities" if it is not accompanied by belligerent nexus. To establish belligerent nexus, it must be determined that the action was purposefully intended to reach the required threshold of harm in one causal step, such that acts of self-defence, the exercise of power or authority, and acts of civil disorder or acts that occur during inter-civilian violence are excluded (Asaro, 2012: 698). There are acts that directly unfavourably impact military operations, or that result in the death or injury of protected persons, which still do not constitute a "direct participation in hostilities" because the perpetrator of the act lacks belligerent nexus. Belligerent nexus requires that over and above the act being causally expected to reach the threshold of harm, it is also necessary that it was deliberate and intentional[103] in order to benefit one party to an armed conflict and to the disadvantage of the

[101] For example, the design, production and deployment of a weapon.
[102] This includes political, financial and social activities.
[103] In the previous chapter I mentioned Asaro's belief that lawful killing must be intentional, but asserted that he was mistaking the philosophical concept with the legal one. Legally, intentional means purposeful, and that is embodied in the concept of belligerent nexus.

other (International Committee of the Red Cross, 2009: 58). Belligerent nexus is not to be confused with the idea of "subjective intent[104]" or with the concept of "hostile intent[105]" (*ibid*. 59). Both subjective and hostile intent relate to the internal mental processes of the subject, while belligerent nexus is an external determination that is constituted by the objective aim of the act (*ibid*.). In other words, wishing another party to armed conflict harm is not enough; the potential target must have acted in a way to bring harm to the other party. For example, the target might wish death upon their opponent, they may even imagine killing the opponent themselves, but it does not constitute belligerent nexus until they begin to act. This distinction is significant for our purposes, since by excluding the need for a subjective investigation into mental states or intentionality, the deeper questions about artificial intelligence and consciousness are avoided. Thus, it would be possible for AWS to meet the requirements of belligerent nexus and so to be considered to be directly participating in hostilities, meaning that it is possible for an AWS to be considered a legitimate target according to the principle of discrimination, especially when viewed with Ohlin's Combatant's stance.[106]

Belligerent nexus is established as a result of the purpose of the act and not as a result of the internal mental state of the individual performing the act. For example, wishing that an enemy gets shot is not sufficient, but going out to buy a gun to shoot an enemy down is. It is an objective criterion and as such is not affected by subjective influences like distress, personal preferences, mental ability, or willingness (*ibid*.). This means that civilians who have been coerced into committing an act that meets the first two categories, and even children who have been illegally conscripted could meet the requirements for "directly participating in hostilities" and may lose protection against direct attack (*ibid*. 60). This is important as it preserves the soldier's right to defend themselves despite the fact that they are potentially facing someone who was coerced into fighting or a child soldier, and it relates to Frowe's conclusion regarding the justness of inflicting harm.[107] As mentioned, Frowe argues that the internal mental states of a target are unknowable, and thus they are not a suitable guide to determining how to act. The consideration of belligerent nexus leads to the conclusion that if an act reasonably appears

[104] A subject's personal state of mind, or their basic motivation behind an action.
[105] A technical term used in the Rules of Engagement referring to the threat of imminent use of force.
[106] The Combatant's stance was discusses in chapter two.
[107] Frowe's distinctions being the justness of inflicting harm and the justness of the harm itself, which were discussed in Chapter 2.

to be designed to cause harm, in lieu of subjective knowledge, there is an objective justification to act, despite the fact that the potential target may be a child or may have been coerced, much like Frowe's conclusion.

The belligerent nexus criterion is designed in such a way that any act that is in self-defence or in the defence of others is necessarily excluded from constituting "direct participation in hostilities", such that civilians are allowed to use force that may reach the required threshold of harm in order to prevent acts that amount to unlawful attack (*ibid*. 61).[108] It is also designed to prevent belligerent nexus being established after the armed hostilities have ceased. Any acts of riots and civil unrest, which do not form a part of the hostilities between parties to armed conflict, fall under the jurisdiction of the civilian authorities and not of the military forces (*ibid*. 62). It also excludes the exercise of power over people or territory by a legitimate authority to facilitate the reestablishment of peace (*ibid*. 63). Acts that form a part of civil disobedience are also excluded, as are all acts that occur during inter-civilian violence, because these are deemed to be under the jurisdiction of the national peacekeeping force (*ibid*.). This would mean that civilians who act in self-defence or in defence of others do not meet the criterion of belligerent nexus, and thus cannot be legitimately be targeted by AWS. Thus, AWS would have to be able to determine if potential targets are acting in self-defence or are just acting in a hostile capacity before being given the authority to make a decision about the application of lethal force. It is tricky to imagine how this would happen, but this is a technical requirement the engineers would have to figure out. If indeed AWS could differentiate in this capacity, then there is no reason to deny them the authority to make decisions about the lethal application of force.

When applying the criteria discussed to AWS, it becomes clear that in order for an AWS to identify a target as a combatant, it would have to be able to identify acts that could unfavourably affect the military's operations, as well as acts that are likely to result in the death or harm of civilians or destruction of their property. More than being able to recognise harm, AWS would also need to be able to monitor actions continually in order to establish a reliable

[108] In armed conflict an unlawful attack could be an act like looting, rape or murder, committed by marauding soldiers.

causal link between the perpetrator of an act and the act that meets the threshold of harm. This means that they would presumably have to be active for a certain duration of time, like a "probation period", before they are allowed to employ lethal force. Lastly, AWS would need to be able to gauge whether the observable behaviour of the perpetrator (performing the action that in one causal step meets the required threshold of harm) indicates that they are acting purposefully. The objective nature of belligerent nexus means that it is established on the basis of observation, but it does require some assumptions to be made.[109] Opponents to the use of AWS would argue that it ought to be unlawful to deploy an AWS that is not sufficiently able to meet these requirements. I would argue that provided the human who is in "meaningful control" does discriminate targets and has at least the ability to override the AWS, there should be no such restriction. In other words, as long as the requirements of IHL are met by at least one agent who initiates, or at least has the ability to override, the decision to employ lethal force, then there is an acceptable causal link for responsibility.

These three elements, namely the threshold of harm, direct causation, and belligerent nexus, are required in conjunction with one another in order to disqualify someone from the protection afforded by IHL. Any application of lethal force that is made without establishing threshold of harm, direct causation, and belligerent nexus constitutes indiscriminate killing, which is a war crime. Any AWS that cannot establish these three criteria, fails to meet the principle of distinction. The use of lethal force by an AWS that fails to discriminate sufficiently between active combatants and civilians would constitute a war crime.[110] Importantly, the above-mentioned factors are guidelines, not rules, and they function to assist a moral agent to navigate complex and varying situations (Asaro, 2012: 698). In this regard, an AWS would have to calculate probabilities and act accordingly, essentially acting just like a human soldier that has to make an educated guess. Both the AWS and the human soldier work with incomplete information, and if it is acceptable for the human soldier to make inferences on incomplete information and act on this basis, then the standard should be the same for AWS. If opponents hold that it is not appropriate for AWS to be allowed to make decisions regarding the

[109] For example, if a person was observed to be armed in an active combat zone, it is generally assumed that they are a part of the hostilities, even though they may not be. Human soldiers make similar assumptions, given that they have no direct access to the internal mental states of potential targets.
[110] In the next chapter I will examine who is responsible for a crime of this nature.

application of lethal force on such a basis, then it is not appropriate for human soldiers either, and the standard for the lawful application of lethal force needs to be adjusted upwards in general. The principle of distinction is not the only IHL principle that needs to be adhered to for an instance of lethal force to be lawful. Even after judgements of distinction are made and a target has been determined to be a legitimate one, there are still limitations to the use of force that is permitted to be used against that target.

4.4. *Jus in Bello* – The Principle of Proportionality

Once a target has been verified as being legitimate in accordance with the principle of distinction, the application of lethal force against that target is permitted; however, any instance of the permissible use of lethal force must further comply with the principle of proportionality (the second principle of *jus in bello*). Failure to comply with the principle of proportionality is also considered a violation of IHL, and the agent that fails to adhere to this principle may be charged with a war crime. As mentioned, IHL is concerned with the treatment of individuals during armed conflict. One of the constraints imposed by IHL is that the use of force not exceed what is strictly necessary in order to neutralise a threat. In other words, the use of force must be proportional to the threat that it is in response to. Proportionality, however, is not as precisely articulated as discrimination, which presents two problems. Firstly, the vagueness of its conceptualisation means that what response is deemed to be appropriate at the time is a subjective judgement made by the actor. This position is too relativistic to be valid guide for acceptable behaviour. Secondly, it presents a significant challenge to those who need to programme AWS to act in a proportional manner. If different engineers have different conceptions about what constitutes a proportional application of force, AWS will too have different frames of reference about what is proportionate. Such differences are important to steer clear of in the law, where it is necessary to delineate what is acceptable and not and to define criminality specifically. If AWS have different parameters for proportionality, it becomes challenging to prosecute anyone for violating IHL on the basis of a lack of proportionality. In other words, it is difficult to prosecute someone for a crime without a precise definition of that particular crime. A discourse on what proportionality entails is necessary if we want to establish whether AWS (and perhaps even human soldiers) are able to adhere to this principle.

The concept of proportionality originally comes from criminal law practices, where it is used as a criterion for fairness and justice to make sure that "the punishment fits the crime", so to speak. Drawing a parallel to international law, the harm inflicted in retaliation to hostilities should be more or less equal to the harm caused by the hostile act. The principle of proportionality is based on the rule against unnecessary suffering or superfluous injury, which is expressed in Article 35(2) of Protocol 1 of the 1977 Additional Protocols to the Geneva Conventions (Anderson and Waxson, 2013: 10). The essence of the principle of proportionality is that any instance of lethal force not be applied in a disproportionate manner when weighed against the gains attained by employing such force.[111] Any disproportionate application of lethal force constitutes a war crime. At this point it is important to note that under the Rome Statute, the death of civilians does not in itself constitute a war crime. It is only when the collateral damage is excessively high in comparison to the direct and concrete military advantage gained that such deaths could constitute a war crime. This means that if there is some military gain to be had from an action, some amount of collateral damage could be justified. The ability to justify collateral damage by saying that the military advantage outweighs it, according to the military's calculation, is why we need to articulate proportionality more clearly; otherwise, this principle is open to misuse. Proportionality should be a more objective standard defined by IHL, rather than something that a particular military has the freedom to define.

As we have seen, the purpose of the provision for *jus in bello* is to prevent someone from being arbitrarily deprived of their right to life and to ensure that any use of force against them is balanced against their right to human dignity.[112] The harm caused to those who act in a manner that constitutes a "direct participation in hostilities", or to the property of these persons, or the collateral damage of a military operation, must be comparative to the benefits to the party that targets them and employs force and must not be excessive when contrasted against these benefits. In other words, it must not be an excessive application of force.

[111] For example, bombing an area where hundreds of men, women and children live in order to secure a faster route between two military bases would arguably be considered disproportionate.
[112] The right to human dignity means that people have inherent value and cannot be treated as a means to an ends.

In some cases, it is easy to see that a party has gone too far in the application of lethal force. For example, the use of "Agent Orange[113]" by the United States against the Northern Vietnamese soldiers during the Vietnam war was undeniably excessive; it was sprayed over large geographical areas killing much of the vegetation, it adversely affected the health of large numbers of the Vietnamese population, and the effects of the poison are still present in the humans whose ancestors were exposed to it two to three generations prior (Aspen Institute, 2016). However, one could argue that the widespread use of poison fails the discrimination principle, so cases like the use of "Agent Orange" do not provide an accurate view of what constitutes disproportionate method of warfare independently from an indiscriminate method. It would be meaningless to define proportionality as anything that is vastly indiscriminate, since this would negate the point of having separate principles. Therefore, proportionality must be constituted by an excessive use of force over and above the lack of discrimination.

There are certain methods of warfare that are so obviously excessive, like biological and chemical warfare, that they have already been expressly prohibited in IHL. The principle of proportionality still exists outside of these obviously excessive methods, so there must be cases of proportionality that are not overtly excessive (because if they were they would presumably already be banned), that also do not overlap with discrimination. If we want to be able to guide AWS in responding proportionately to an attack, and perhaps if we want to provide a clear guide to proportionate responses to human soldiers, we ought to articulate the principle of proportionality more clearly.

The question of what exactly constitutes an excessive application of force remains. Human Rights Watch compares "excessive" applications of force to "arbitrary" applications, since the latter concept has been more expressly articulated (2014: 9). Arbitrary killing is more specifically defined in the 1990 Basic Principles on the Use of Force and Firearms guideline, where the United Nations formulated the guide for law enforcement officers regarding the use of firearms (*ibid*.). While this document is aimed at regulating the use of firearms, it is one of the few instances of international legislation where there was an attempt to qualify the

[113] A herbicidal dioxin containment poison.

appropriate use force more precisely. Thus I will recount the conditions for the appropriate use of force presented in this document, and endeavour to apply it to proportionality in the context of the use of AWS in warfare. In the Basic Principles on the Use of Force and Firearms it is specified that the non-arbitrary application of force requires adherence to three conditions: (1) the force must be necessary, (2) it must constitute a last resort, and (3) it must applied in a proportionate manner (*ibid*.). In the discussion of these conditions, it is apparent that they overlap in many ways. Furthermore, it may seem strange to define the principle of proportionality with the sub-clause that force must be applied in a propionate manner, but as is shown, the Basic Principles on the Use of Force and Firearms articulates the latter criterion in a more utilitarian manner than an understanding based on fairness.

The requirement of necessity states that members of law enforcement agencies are permitted to use lethal force only in situations where it is "strictly necessary" to prevent the loss of life, for example, in circumstances that qualify as instances of acting in self-defence or in the defence of others (*ibid*. 10). Even after a target has been identified as directly participating in hostilities, it is still not permissible to employ lethal force against that target without determining that the application of lethal force will save lives, such that if the harm caused by the target will cause only minor injury or the destruction of property, lethal force is not a proportionate response. This condition correlates with that mentioned in the discussion on *jus ad bellum* – the government would be allowed to authorise the use of lethal force (or of AWS with lethal capacities) if doing so would save lives immediately. If an AWS is to respond proportionally to a threat, it should be able to recognise instances of mortal peril and only employ lethal force against people "directly participating in hostilities" if it would save the lives of military personnel or civilians. This requirement is not sufficient to define proportionate response on its own, as it is too similar to the "threshold of harm" criterion in the principle of discrimination, thus there must be more to proportionality than reactions relating to "self-defence" or "the defence of others".

In addition to the force being necessary, it should also constitute a last resort. In order to constitute a last resort, all other alternatives must be exhausted or "without promise of achieving the intended result", meaning that alternative, non-lethal methods would be fruitless

(*ibid*. 11). In light of this requirement, law enforcement officials are trained in "methods of persuasion, negotiation and mediation", so that they have the means to bring about peaceful resolutions to hostile situations (*ibid*.). In order to make it possible for AWS to fulfil this condition, the AWS may need to be equipped with alternatives to lethal force. For example, they may need to be equipped with non-lethal weapons[114] and may even need to be endowed with the ability to "talk down" hostiles. Otherwise, AWS may need to have restricted autonomy, such that they cannot employ lethal force without approval from a human operator with meaningful control over them, who needs to be the one able to employ non-lethal force and negotiate peaceful resolutions. The problem with this solution is that the system would not qualify as an AWS according to the definition established in Chapter 3. This means that these systems are essentially not AWS, and are bound by the previous existing legal system. This solution seems tempting, but it is not feasible in the long run, as it does not cater for the kinds of fully autonomous AWS that will one day exist. Thus, it is irrelevant for the purposes of this discussion. The best solution is to equip lethal AWS with non-lethal weapons as well, which they can be programmed to employ before applying lethal force, and lethal force can be restricted to be only being applied when it saves lives immediately. This would meet the requirement of "exhausting possible alternatives" and caters for the inevitable development and use of fully autonomous AWS, as defined in Chapter 3.[115]

The last requirement is that of proportionality, namely, law enforcement is required to "act in proportion to the seriousness of the offence and the legitimate objective to be achieved"; they are obliged to act conservatively in their decision to employ lethal force and to always minimise the harm they cause (*ibid*. 12). It seems circular to say that proportionate applications of force should be proportionate, but I will attempt to unpack this requirement further to show that within the principle of proportionality, there is a utilitarian criterion.[116] Usually, in order to conform to this condition, law enforcement officers assign values to the targets, objects, and

[114] From simpler non-lethal weapons like pepper spray and Taser guns, to more technologically complex ones like long range acoustic devices, which make the victims ears ring uncomfortably loudly, or active denial systems, which target heat rays such that the victim experiences the sensation of burning without physical damage.

[115] Specifically, that autonomy in AWS is qualified by the fact that they are able to decide to employ lethal force without awaiting approval from a human operator.

[116] This utility calculation is narrower than the general principle of proportionality, which additionally requires that force necessary and a last resort.

even to the human beings that are involved in, or proximate to, the application of force, much like a utility calculation (Anderson and Waxson, 2013: 13). This is also the case in military operations, and human soldiers make these utility calculations as well. In order to conform to this principle, AWS would have to make these kinds of calculations as well. Assigning numerical values to human life appears somewhat arbitrary and a seemingly impossible task, but law enforcement officials and soldiers are trained to make these kinds of utility calculations, so there must be methods already devised to guide an agent to make these assignments. If we can train people to do something so seemingly counterintuitive, we could theoretically teach a machine to do so as well; similar to assigning values of importance to different chess pieces. Thus, it is conceivable that an AWS would be able to perform the same kind of utilitarian decision making in order to conform to this condition.

Machines could learn proportionality from humans; initially the human operator who has meaningful control over the AWS (or a trainer) might be the one to make these calculations. The AWS could learn from the human operator's patterns of decision making. Before AWS are allowed to actually make these decisions, we could allow them to provide the conclusions they think adhere to these principles and then correct them. This is technically how crude mechanisms of machine learning occur. This testing could happen over a conservative period of time. As with the principle of distinction, the criteria discussed above are formulated in a purposefully vague manner in order to allow for basic human judgement and decision making. This would mean that it is impossible to pre-programme AWS with the correct responses to moral dilemmas, and AWS will have to make the best decisions they can subject to the information they have, like human soldiers do. And like human soldiers, their decisions should be subject to review, either by the commanding officer or by other official review boards.[117]

Up until this point I have discussed the current legal parameters for the permissible use of lethal force, and examined some ways in which AWS could be developed and deployed in order to conform to IHL. However, there are not yet legal guidelines for AWS specifically. This is an issue that needs to be addressed; if there is no legislation regulating AWS specifically,

[117] It is standard practice to review the success of an operation in order to make adjustments in the future, and it stands to reason that similar review practices will be instituted with AWS.

it becomes easier for the misuse of AWS and associated war crimes to go unpunished. For this reason, I will now discuss the transformation of the existing legal structures to cater for the use of AWS.

4.5. Regulating Autonomous Weapons Systems

The existing legal framework of IHL has yet to deal with the relatively new issue of autonomous weapons. If ever there was a motivation to change the way we interpret IHL, it is the potential of robots becoming agents in our world, which challenges the assumption built into the law that only humans can act as agents (Asaro, 2016: 2). It is important to remember that any law, including IHL, is not just a single piece of legislation and that there are many components to it, including the applications of current legal customs, which may not be codified, and case law, which is also constantly being updated (*ibid*. 3). Thus there is not only one document that needs to be altered, nor will the promulgation of one piece of legislation address the issues raised by AWS. Many existing pieces of legislation need to be reinterpreted, and the codification of any newly emerging norms will take time. This reiterates the idea that the discussion of the acceptable use of AWS needs to happen now; delays in this process means a bigger legislative gap for people to misuse AWS and for the associated war crimes to go unpunished.

There are four possibilities for the transformation of existing norms in light of the introduction of new technology. First, the existing norms might function adequately enough to continue to govern practices relating to the use of new technology. Second, the existing norms might need to be diminished or dissolved. Third, new norms may emerge to govern practices relating to the use of new technology. Lastly, the existing norms could be adapted or reimagined in light of the new technological development. In this last case, while the fundamentals of the law remain more or less stable, interpretations of specific pieces of legislation can change (*ibid*. 1). The use of AWS is currently restricted, but in part only due to regulations the US DoD has enforced on its self. While IHL currently requires discriminate and proportionate use, these criteria are vague and open to interpretation. Thus for the most part, new legislation will need to be drafted to ensure meaningful and lasting arms control.

There is a need to decide which norms to codify into legislation. Currently, the norm is emerging that a human must be involved at some stage in the decision to apply lethal force, or at minimum that a human must be able to override the AWS. However, even if consensus is reached that a human must be proximate to the decision to employ lethal force, the question arises as to how such a norm will be established in practice (O'Connell, 2014: 232). One way of establishing such a norm is to employ strategies previously used to control weapons we already have, including outright bans on certain weapons, restrictions on how and where to use them, and limits on who may use them (*ibid*. 233). The core sentiment of this norm is opposition to removing humans from the "kill chain" (*ibid*.). The requirement that a human being makes, or is at least involved in, the judgment to lethally engage a target is a means of enforcing arms control, but this norm has yet be formalised in IHL (*ibid*.). In international law the most common way to build new norms is through multilateral treaty negotiations, which means that governments need to meet and come to an agreement about the use of AWS (*ibid*.). Bureaucracy moves slowly, so these treaty negotiations will likely lengthen the arms control process. In the meantime, we need to figure out how to expedite this process. In this regard, the ICRC has put forth a number of proposals to guide regulations regarding AWS to ensure that their use is compliant with IHL.

There are a number of practices that states can unilaterally implement relatively quickly in order to enforce arms control regulation. These regulations may not form a part of IHL until they become a part of a multilateral agreement or are codified formally in another capacity, but having governments address the use of AWS could serve as a short-term solution. The ICRC recently published a document where they offer potential state-implemented measures of arms control over AWS. They suggest that both the predictability and the reliability of the weapon in question need to be dependably established, such that the human who deploys the AWS can make reasonable assumptions about the consequences of its use (International Committee of the Red Cross, 2016: 5). Furthermore, the capacity for human intervention needs to be clear before the weapon is deployed, and, in that spirit, accurate information regarding the AWS in question needs to be readily available to all parties involved in their use and regulation (*ibid*.). They also urge that all issues regarding accountability are resolved before any AWS are deployed (*ibid*.). I will return to issues of responsibility in the next chapter.

Arms control measures take time, and in the interim, the ICRC proposes a number of strategies states can adopt in lieu of international consensus. Many of the suggestions that the ICRC puts forth are based on current practices to restrict existing weapons that possess some level of autonomy (*ibid*. 2). One way to limit autonomy is by imposing operational constraints, for example, by limiting the tasks an AWS is permitted to carry out, perhaps even reducing its operations to a single function. For example, the AWS could be imbedded with the singular directive to stop incoming enemy projectiles. This would make the behaviour of the AWS more predictable (since the scope of behaviour is narrowed), which means that the AWS will be easier to control. They also suggest that one could limit the potential targets an AWS may engage, possibly to objects or vehicles, and leave human targets out of it for the time being. AWS can further be constrained with limitations on the operational environment they are allowed to engage in, such as restricting their field to a geographical area. Lastly, the ICRC advocates including procedures for human intervention in AWS attacks, whereby a human can deactivate the weapon (manual override). The arms controls measures proposed by the ICRC seem like reasonable constraints on behaviour, similar to the rules human soldiers have to abide by. As such, it would not render the weapons as merely being automated or passively-autonomous, and would not defeat the purpose of having an autonomous weapon to begin with.

Ultimately, the goal would be to have arms control measures that are internationally recognised and apply universally; in particular, ratifications to the proportionate use of force is necessary. But these measures will take years, if not decades, to institute on a global scale. The hope that the ICRC has is that states will take responsibility for the development of AWS and will impose legal restrictions on their own militaries. This may seem unlikely, but such measures may come to pass if the use of AWS becomes an issue in political campaigns. It is however, unwise to put too much faith in this avenue of arms control. The problem is that without IHL ratifications, even if states take on the responsibility of arms control, there is no lasting standard for the permissible use of AWS. Arms control treaties that are multilateral and global will take time to implement, but once they do become law it will be significantly more difficult for countries to disregard the rules for the lawful use of AWS. Nonetheless, the strategies proposed by the ICRC are good examples of the types of arms control measures that could be codified in IHL that do not render the weapon system as being passively autonomous, thus they are worth noting.

4.6. Summary

One of the first obstacles advocates of AWS will have to overcome is to clarify how AWS will be able to satisfy the conditions for the permissible use of lethal force imposed by IHL. Lawful killing must be done in a way that humanity, specifically with regard to the right to life and human dignity, is preserved. In the examination of IHL it appears that the standards that need to be adhered to do not hinge on human judgement itself, rather it appears that human-like judgment is sufficient, meaning the ability to reason whether the circumstances match those described in IHL.

The practical application of the theoretical principles codified in IHL is a complex task and its just and effective completion calls for interpretation of the relevant laws at every turn. IHL explicitly obliges agents to consider the repercussions of the decisions and actions they take, and rests on an appeal to their humanity, calling for these decisions and actions to reflect empathy and integrity (Asaro, 2012: 700). We must be aware that the deliberations on the lawful use of lethal force occur on the basis of vague, incomplete and evolving principles, which at times function in opposition to one another, and attempts to draw insight from this confusion (*ibid*.). Articulating legal requirements into a syntactic formula is a gargantuan task, and the sooner policy makers embark on this endeavour, the better. Arms control takes time, and we do not want to end up with a gap between the use of AWS and their regulation.

The Martens Clause is often cited as a reason for banning weapons in disarmament conversations because it provides a sort of “last resort” defence to people, since it essentially states that just because something is not expressly prohibited does not make it permissible. In this way, the Martens Clause covers the gap between the development of a new weapon and the regulation of that weapon. In addition, it embodies the spirit of IHL since it requires that parties to armed conflict respect the “principles of humanity” and the “dictates of public conscience”, even if the two concepts are not explicitly defined. These concepts can broadly be understood to mean that parties to an armed conflict should respect the right to life and human dignity and should act on the conservative side when deciding to apply lethal force. For AWS, this means that just because there are no specific regulations against their use, it does

not mean that their use in an unconstrained manner is permissible. Their use would have to adhere to the basic principles of humanity and the dictates of public conscience. But both of these concepts are vague and subject to interpretation, as such it would not be prudent to rely too heavily on the Marten's clause to curb the use of AWS. Thus, we must examine other areas of the law if we are to forge meaningful arms control measures.

Currently, the norm that AWS ought only to be used in armed conflict (like cruise missiles) is building. The decision on whether or not AWS could be deployed would then be governed by *jus ad bellum*. For technical purposes, AWS are as old as any weapon[118] if they are defined as a weapon that may be triggered by the target rather than by the user (O'Connell, 2014: 224). For purposes of *jus ad bellum*, this comparison between AWS and older, less autonomous weapons, is enough and no significant changes are necessary in order for AWS to be considered as a viable weapon in a just war. The guidelines provided by *jus ad bellum* detail situations where government may sanction the use of AWS, the most obvious and critical use being to save lives immediately. The only task that needs to be completed here is to decide if an AWS is indiscriminate enough to be restricted to only being used in armed conflict, like cruise missiles, but this judgement will have to be made on a case by case basis, since not all AWS are inherently indiscriminate. Whether or not AWS are restricted to use in armed conflict, their use will also be governed by *jus in bello* and the principles of distinction and proportionality.

The two aspects of *jus in bello* that are critical for an AWS to comply with are the principle of distinction and the principle of proportionality. The principle of distinction requires that an AWS validate the legitimacy of a target by establishing that the target is a direct participant in hostilities, on the basis that their actions will directly and causally result in a certain threshold of harm and that the act was orchestrated with belligerent nexus. Human soldiers make this distinction based on observation and inductive probabilistic reasoning, which is based on the recognition of patterns. AWS would have to have the same probabilistic reasoning capacities, in order to be able to distinguish as effectively as a human soldier. In this

[118] From primitive booby-traps to antipersonnel landmines

regard, only slight changes have to be made to the existing of laws, such that the principle must be satisfied by either the human operator who has meaningful control over the AWS or by a fully autonomous weapons system with the capacity to sufficiently discriminate between combatants and civilians.

The greatest obstacle lies with the principle of proportionality, which is less clearly defined, and this is another ambiguity that needs to be clarified in order to enact meaningful arms control. From examining IHL, it seems that the disproportionate application of force is constituted by collateral damage that is unacceptably high. This is essentially the same as a vastly indiscriminate application of force. There must be more to proportionality that differentiates it from discrimination to justify having the principle of proportionality in the first place. Force that is not proportionate could also be deemed excessive, but cases of the use of force that are obviously excessive tend to be prohibited in any case, so there must be non-obvious and still discriminate applications of force that necessitates the articulation of this principle. Human Rights Watch uses the concept of arbitrariness in order to try to expound on what conditions could constitute the proportional use of lethal force. Specifically, the non-arbitrary use of lethal force is necessary and constitutes a last resort, and is proportionate in a utilitarian sense. That is to say, such lethal force is necessary for self-defence, or because other, non-lethal methods will be ineffective, and it should not be in excess of the military advantage gained (according to a utilitarian calculation). Human soldiers are presumably (hopefully) schooled in what a proportionate response is, so, theoretically, an AWS could also be. At the very least, this understanding provides a starting point by which to understand proportionality, which will gives programmers of AWS something to work with.

One possibility to satisfy the abovementioned legal principles is to adopt Ohlin's Combatants' stance.[119] With this viewpoint, the circumstances in which an AWS can use lethal force should not be different to when a human is allowed to kill, as discussed above. The above-mentioned legal guidelines, most importantly, the principle of distinction and the principle of proportionality, must be adhered to for killing to be lawful. These principles must be adhered

[119] The Combatant's stance was discussed in Chapter 3.

to by the machine. However, if the machine lacks the capacity to comply with IHL, it is possible to view the operator as the agent obligated to ensure IHL compliance. Repeating Asaro's point on meaningful control[120], part of the reason for keeping a human in- or on-the-loop is to ensure compliance with IHL, and to provide an appropriate locus for responsibility.[121] Thus, the human operator is one who must ensure that the above-mentioned principles are adhered to. As mentioned in the previous chapter, if the use of a human supervisor is to be considered a legitimate means of arms control, it must be formalised in legislation.

Ultimately the developers and users of AWS will have to contend with all the legal requirements outlined in this chapter. AWS will have to be able to determine whether targets are friendly or hostile (to comply with the principle of distinction), and whether the current battle conditions meet the ROE for weapons release, and if so what the chances are of collateral damage, amongst other things. Like humans, they will have to make their best guess relating to all of these issues (Cummings, 2014: 6). It stands to reason that if the semi-autonomous weapons systems currently in use can establish that an attack is imminent and react accordingly[122], then fully autonomous ones ought to at least be able to perform as well, thus it is conceivable that AWS could perform in the abovementioned manner. Given the discussion in this chapter, it is likely that amendments to the current legal structure will have to be made, and that will be time consuming. In the meantime, the existing mechanisms in international criminal law are currently amenable to prosecuting an AWS that perpetrates a war crime. This point will be made clear in the next chapter, which focuses on responsibility. For now, the law functions well enough, though barely, to establish scenarios when AWS is permitted to engage in lethal force. Legal ratifications still need to be made to eliminate the ambiguity of proportionality. However, if AWS can function similarly enough to humans in discriminating targets and determining a proportionate response, and it is conceivable that they will be able to do so, then they are legally permitted to take a life. What remains unclear is who is responsible

[120] Asaro's full argument on meaningful control was discussed in Chapter 3.
[121] The next chapter discusses the appropriate allocation of responsibility.
[122] Several weapons systems with varying levels of autonomy are already in use in a defensive capacity, including the Counter Rocket, Artillery and Mortar system. The C-RAM is able to detect incoming strikes and intercept them (globalsecurity.org).

when these conditions were not sufficiently satisfied, which is where my discussion turns to now.

5. Responsibility

Establishing when lethal force is permissible is only one of the challenges facing the use of AWS. Even after this determination has been made, another critical issue remains. The second chief concern opponents have, and an important one to address before AWS are deployed, is over where responsibility lies when a machine employs lethal force in an unlawful manner. In normal cases of criminal culpability, it is necessary to establish *mens rea*[123], or malicious intent, in order to assign blame (Dennett, 1997: 352). Additionally, in military contexts, there is normally a chain of command where the commanding officer takes the responsibility for the actions of his subordinates and liability is based on mechanisms of *respondeat superior* (Asaro, 2011: 178).[124] However, when one considers that autonomous weapons systems are designed and programmed by one group of people and deployed by another, and when one considers the possibility of more than one AWS cooperating in an attack, assigning responsibility for an action that violates one of the principles of IHL becomes a more complicated task. Even more so, as the AI develops the ability to act of its own accord as artificial reasoning and decision making evolve, the distinction between the *a priori* protocols programmed and the commands given by the commanding officer, and eventually its own decisions, will disappear (*ibid*. 175).

Deborah Johnson, a professor of applied ethics, claims that if one examines the relationship between humans and technology it is obvious that people are responsible for how technology is used (2014: 2). This is true in most cases, but in the case of an autonomous artificial agent with the capacity to learn as it operates, it is difficult to uphold this claim. As the agents become increasingly autonomous they become artificial moral agents, and no human will be responsible for their behaviour (*ibid*.). Thus there is a gap in responsibility between the programmers, the officers who deploy the weapon, and the autonomous system itself, which

[123] The idea is that an act is not culpable unless the mind is guilty. This means that the agent must knowingly have intended the unlawful action.
[124] A Latin phrase that means "Let the master answer". In this context, the master would be the commanding officer.

Andreas Matthias calls the "Responsibility Gap" (*ibid*.). The Responsibility Gap is one that must be bridged before AWS can be legally and ethically deployed.

While it is challenging to assign blame to individual persons in this case, in most circumstances we do want humans to bear some degree of accountability, otherwise AWS could be used in a manner that either contravenes IHL, or at least violates the spirit of its principles, without any repercussions. The objection over the lack of accountability is based on the idea that it is legally necessary to maintain direct individual human responsibility in all cases of the use of lethal force, including any instance of lethal force employed by an AWS, despite the fact that the person who bears responsibility may not be in direct control of the behaviour of the AWS in question (Allen and Wallach, 2013: 126). However, there is no requirement to maintain individual, human liability in the law. Furthermore, as will be shown in this chapter, Jens Ohlin (2016) and Peter Asaro (2011) point out existing mechanisms for assigning blame that can easily be adapted to suit AWS, which means that it may be possible to hold someone other than the person who deployed the AWS responsible.

In the event that an AWS violates one of the principles of IHL or commits an act deemed a war crime, Robert Sparrow (2007, 69-73) identifies three likely candidates for responsibility: the programmers, the commanding officer, and the machine itself. He argues that it is not possible to meaningfully hold any of these candidates responsible and uses this as an argument against the use of AWS. I will examine Robert Sparrow's argument, supplemented by Jens Ohlin's discussion of the existing legal precedent for assigning blame regarding commander responsibility, by Allen and Wallach's paper on arms control and artificial moral agents, and by Daniel Dennett's explanation of intentionality in AI regarding the possibility of machine liability. I will use all these theorists' views to reject Sparrow's claim that there are no good candidates for assigning responsibility to, so too rejecting the so-called "responsibility gap". I will also offer other possibilities for assigning responsibility, in order to show that there are already suitable methods for assigning responsibility, and thus that this issue is not as significant a challenge as opponents would argue.

5.1. The Programmers

Robert Sparrow is a professor of ethics at Monash University, and is a well cited author in the discussion of AWS. He starts his pursuit for the locus of responsibility with the logical first point of origin of AWS; the programmers (2007, 69).[125] It seems intuitive that the creators of a weapon ought to bear responsibility for its use; after all, they built it with a lethal purpose in mind. The issue here is with autonomy itself; if it is acknowledged that the AWS can choose for itself, it is possible that the AWS will at some point select a target other than one it was originally programmed to select, and this makes the weapon "unpredictable". This means that the programmers could not possibly have the necessary mental intention to be convicted for a war crime. There are existing regulations against the use of unpredictable weapons, since they generally fail on the principle of discrimination. Sparrow, however, counters the assumption that the programmers ought to bear all responsibility for war crimes committed by AWS by saying that the possibility that the system could identify or engage the wrong target is an acknowledged limitation to AWS (2007: 69). In other words, Sparrow claims that despite the unpredictable nature of AWS, their unpredictability is not erratic enough for the outcomes of their use to be completely arbitrary. This means that AWS will generally perform in a manner that is broadly constituent with what they are designed for, and will only sometimes perform in an unexpected manner. The inherent unpredictability in AWS with low levels autonomy is not higher than the uncertainty associated with other weapons systems, and the unpredictability in AWS with high levels of autonomy is not higher the uncertainty associated with people. Therefore, it is not possible to hold the programmers responsible for a war crime based on recklessness either, as they have not intentionality created an intrinsically unpredictable weapon. Sparrow argues, and I agree, that it is only fair to place the burden of responsibility on the programmers if the programmers where careless or reckless in the engineering of the system and as a result of that carelessness the system violated some principle of IHL or perpetrated a war crime (*ibid*.). This is consistent with current responsibility practices. Deborah Johnson agrees and suggests that the programmers should only be held responsible in the same manner as engineers, namely on the grounds of professional responsibility (2014: 3), which echoes a proposition put forward by Peter Asaro.

[125] I shall speak only of the programmers but this discussion is relevant to the designers, engineers and manufactures; all these titles are equivalent for my purposes.

Asaro suggests that existing product liability law could be used to assign responsibility to programmers. In fact, he believes that it is already applicable to AWS, since these laws cover all manufacturing (2011: 171). Under these laws, if the manufacturer made a product that produced unintended consequences often enough, they may be required to pay damages under civil law, and if they were negligent in the process, they could also be held criminally liable (*ibid*.). In normal cases, assigning legal liability to the manufacturer based on negligence rests either on the manufacturer's failure to warn consumers about the potential for deviance, or on their failure to take proper care in assessing potential risks (*ibid*.). The mechanisms of product liability law are essentially what Sparrow is suggesting we use in order to hold programmers responsible in cases of negligence. The programmers would presumably engineer the AWS to have a certain scope of behaviour[126] that it normally functions within, and if it functions outside of that intended scope too often as a resulting of engineering flaws, the programmers could be held responsible. The only problem here is that "proper care" is normally defined according to standard industry practices, and since this is a relatively new industry, it may prove trickier to assign blame based on deviation from industry standards (*ibid*.). This trickiness is compounded by the fact that AWS operate in highly variable environments and learn while they do so (*ibid*.). Asaro suggests that the development of "industry standards" is an important topic in roboethics for the future (*ibid*. 172). Asaro further argues that if the operator uses the machine outside of recommended use[127], then they are liable and not the manufacturer (*ibid*. 174).

The abovementioned suggestions are common practices in liability law, and legal action on this basis should not require any special changes or re-formulations of legislation to be effective. If the programmers made the unpredictable nature of the weapon apparent to those who purchase and deploy them, the manufacturer can no longer be held responsible[128] for such unpredictability (Sparrow, 2007: 69). In these cases the responsibility for any unpredictable actions logically shifts to the individual or individuals that elected to activate and employ the AWS in spite of their unpredictability, much like people who choose to drive cars accept responsibility that they may lose control over the car and cause an accident. The purchasers of

[126] As with any manufactured product, it is likely some leeway will exist to allow for a small degree of unpredictability.
[127] For example, if an AWS's sensor technology is unreliable in rain, the recommended use would be to not deploy the weapon when it is raining.
[128] This is a standard indemnity clause which is used often in practice.

the weapons would have to know that there is a possibility that, given its autonomy, the AI is able to and will inevitably make decisions that are not the decisions embedded in and encouraged by its software; after all, the freedom to choose how to act is intrinsic to the definition of autonomy (*ibid*. 70). Sparrow then concludes that as the autonomy of the system increases, it would not be possible to hold the programmers responsible; any unpredictability that occurs in instances where the programmers where not negligent is normalised deviance[129] (*ibid*.). Sparrow's position that only *mala fide*[130] or negligent programmers ought to be held responsible is consistent with existing product liability practices, so I am happy to agree with him on this point. We have also seen that it is already possible to hold negligent or reckless programmers criminally culpable according to product liability law, so this issue does not require further deliberation.

5.2. The Commanding Officer

If the programmers cannot be meaningfully held accountable for the actions of AWS, the next logical candidate for responsibility must be the person or persons who choose to deploy the weapon, specifically the commanding officer (*ibid*.).[131] Indeed, this seems consistent both with current military practices and with the US DoD's directive on the future use of autonomous weapons. In Chapter 3, I have also argued that the human operator who has "meaningful control" over the AWS should ensure compliance with the principles of IHL and, to some degree, accepts responsibility for their failure to ensure this compliance. The commanding officer would be well aware of the risk that the system may go awry and this risk is accepted in the decision to deploy the weapon (*ibid*.).

Sparrow, further, argues against holding the commanding officer responsible. He begins his argument by saying that if weapons autonomy is understood in such a way that its autonomy necessarily implies that at times it will perform unreliably and unpredictably, which means that unintended fatalities will be unavoidable, then the unpredictable nature of AWS is

[129] This refers to the process where normally unacceptable results become acceptable.
[130] Acting in bad faith or with intention to deceive.
[131] The commanding officer is the human operator who has "meaningful control" over the AWS. This discussion is also applicable to any authority that sanctions deployment.

equivalent to the weapons we currently utilise, and no further discussion on commander responsibility is warranted, since the existing practices suffice (*ibid*.). However, Sparrow believes this assumption is incorrect, as it regards "smart" weapons as comparable to "dumb" weapons and denotes no essential differences between the two classes (*ibid*.). I agree with Sparrow that this is not true; there is a fundamental difference, namely that "smart" weapons can make their own choices in target selection and the use of lethal force (*ibid*.). This autonomy implies that the commanding officer cannot fairly be held responsible for the actions of an AWS any more than he could be for the actions of human soldiers under his command.

Sparrow further claims that the more autonomous the weapon, the more likely it is that there will be deviances from command, and thus the less assurance we can have that AWS will perform as anticipated (*ibid*.). This undermines any notion of meaningful control, as control is qualified by predictability. For Sparrow, this means that the commanding officer could not have meaningful control and therefore could not be held accountable for the actions of AWS. In addition, he believes that the faith that some people have in AWS' potentially superior capacity to differentiate between combatants and civilians, and the subsequent claim that they are morally superior weapons[132], is unfounded. Sparrow says an unpredictable weapon characteristically cannot be morally superior. He writes that part of the unpredictability of AWS entails that the AWS can choose to disobey commands, and so the orders the AWS receives do not necessarily determine what action it will perform and what the outcome can be expected to be (*ibid*. 71). This reiterates his conclusion that the more autonomous the weapon, the less fair it would be to hold the commanding officer responsible.

I disagree with Sparrow's point here, because any human soldier has at least the same degree of autonomy. I believe accepting Sparrow's claim culminates in one of two outcomes: either the AWS has too much autonomy to perform reliably and as intended, and therefore human soldiers also have too much autonomy to perform reliably and as intended, or AWS can perform reliably despite their autonomy, just like human soldiers are expected to do. Thus Sparrow's assertion that the autonomy of AWS definitely absolves the commanding officer of

[132] This is notably claimed by Arkin, as discussed in Chapter 3.

responsibility is unfounded. I do agree that we should not categorically allocate blame to the commanding officer, since there are circumstances where the programmers or other parties will be the appropriate party to blame, but there are some circumstances where the commanding officer ought to be held accountable and where assigning responsibility to him or her is appropriate. I will further elaborate on this point by discussing related arguments presented by Peter Asaro and Jens Ohlin.

Asaro (2011: 176) first establishes that there are multiple cases in the law where an independent autonomous agent is acting on behalf of, or on the orders of, another agent and where there are established mechanisms for assigning blame to the other agent.[133] Asaro's point undermines Sparrow's claim that the commanding officer cannot be held responsible for the actions of an AWS. He also points out that many scholars have made a legal comparison between AWS and domesticated animals, which seems to adequately address the issue of responsibility (*ibid*.). Like domesticated animals, AWS have some autonomous capacity, but they are legally regarded as lacking moral agency and culpability. In addition, they are not dangerous if they are "kept" properly (*ibid*.). Asaro goes on to argue that in certain situations the owners of AWS, just like the owners of animals, would be liable for damages caused even if the owner themselves lack culpability (*ibid*.). To illustrate his point, if a dog bites someone, the owner is liable for the medical damages even if he or she did not intend for the dog to bite the person. Furthermore, in the case of animals, there is never an inquiry into the intentionality[134] of the animal; it is only the owner's intention that is relevant (*ibid*.).[135] This helps us avoid tricky questions about machine intentionality. Owners are liable for damages caused even if they lack culpability. Furthermore, if owners are negligent, reckless, or even purposeful, they are criminally liable (*ibid*.).[136]

Asaro suggests that if an AWS, like an animal, suddenly acts erratically or unexpectedly, or disobeys its owner, the owner is usually considered to have diminished liability, though not

[133] Examples of such cases include employees or soldiers following the orders of their superiors (*ibid*.).
[134] Intentionality was fully discussed in Chapter 2.
[135] This is not to say that the owner needs to have intended the criminal outcome, but only that they were aware that in owning an animal that could act in a certain way, they would be responsible for it.
[136] I will discuss recklessness in more detail shortly, under Ohlin's argument.

necessarily no liability whatsoever (*ibid*. 177). Furthermore, in the case of dangerous animals, there are special laws that govern their ownership and the very act of owning them knowing they are dangerous places some liability on the owner (*ibid*.). The same would apply to AWS. Asaro goes on to point out that there is already technology that is considered dangerous if used improperly (cars, planes, and guns, for example), and therefore their use is regulated (*ibid*.). AWS, like other weapons systems, would be considered dangerous if used improperly, thus their use will also be strictly regulated. He argues that treating robots like animals easily solves the problem of culpability without requiring new legislation be drafted (*ibid*. 178). I agree with the legal analogy offered by Asaro, but I believe there is further work to be done in differentiating levels of culpability and acts that are war crimes. Such distinctions in culpability are made in domestic criminal law, and should apply in international criminal law as well. In IHL, for example, one could differentiate between a commanding officer who knowingly endangered civilians and one who unknowingly did so; the former is clearly guilty of a serious crime while the later may be guilty of a lesser offence or potentially not guilty at all. The necessity for such distinctions will become obvious. In this regard, Ohlin offers an argument similar to Asaro's but in a context specific to IHL.

Jens Ohlin, an acclaimed legal scholar, believes that there are existing mechanisms in international criminal law that can be adopted to include AWS and relieve some of the unease opponents feel regarding issues of responsibility. He uses the Nuremberg trials and the Eichmann case as examples of a legal apparatus that can be reframed to this end (2016: 1). Firstly, he considers the basic framework for AWS liability to come from the Nuremberg trials.[137] The aim of Nuremberg was to pave the way for individuals to be held liable for war crimes, even if they only indirectly contributed to the crimes in question, regardless of whether the instrument of their criminality was also a morally culpable agent or not (*ibid*.). Existing international criminal law was reframed to include modes of liability for convicting those who indirectly contribute to war crimes through a "machine" or organised "apparatus" of power (*ibid*.). The use of the term "machine" here is meant metaphorically, to indicate individuals

[137] Trials after World War Two where the Allied forces presided over the trials of prominent members of the Nazi party and their war crimes.

who form together in a "machine-like" organization, but Ohlin believes that the technical requirements of this terminology apply to an actual machine like an AWS (*ibid*.).

As discussed in Chapter 2, Ohlin approaches sophisticated AWS with what he refers to as the "combatant's stance[138]", which is the position that one would have to take in order to make sense of the decisions and actions of an AWS. Adopting this position means attributing aspects of cognition to machines on the basis of human-like behaviour (2016: 2). As will be shown, Ohlin believes that if AWS are viewed with the Combatant's stance it would be possible to hold the commanding officer accountable for the commission of a war crime executed by an AWS without determining whether the AWS in question is a culpable agent or not (*ibid*. 2).

Ohlin uses the case of the Borkum Island war crimes trial to present the idea of collective action[139] as a basis for assigning responsibility to those that indirectly contribute to a war crime (*ibid*. 3). During World War Two, seven American pilots crashed on Borkum Island in Germany and subsequently surrendered to the German soldiers (*ibid*. 4). According to the existing Laws of War, these soldiers ought to have been considered prisoners of war and sent to a camp for prisoners of war, however, they were marched through the streets of the town, and the citizens were incited by Nazi soldiers to assault them, which culminated in the soldiers being executed (*ibid*.). During the trial, the prosecutor was unable to assign individual blame for particular results, since so many individuals had participated, and the aggregation of these individual acts of criminality combined together to produce a horrific outcome (*ibid*.). As a result, many individual persons were tried together on the basis of "collective liability". Ohlin believes the concept of "collective liability" can be used to link the actions of AWS to the orders of the commanding officer to establish a causal link for responsibility, such that the AWS and the commanding officer are viewed as acting collectively.

[138] This concept was fully explained in Chapter 2.
[139] Multiple individuals might combine together to form a group agent capable of achieving a common objective.

Similar war crimes committed in the former Yugoslavia led to the establishment of the Joint Criminal Enterprise (JCE) legal doctrine and the formalization of the concept of "collective liability" (*ibid*.). The purpose of the JCE was to facilitate the conviction of defendants for their participation in a machine-like enterprise even if they were not the physical perpetrators of the act (*ibid*.). The idea is that the orchestrators, who facilitate acts but do not participate directly, are just as liable as the "cogs" for the criminality the machine produces (*ibid*. 5). A key concept of the JCE doctrine is *organisationsherrschaft*, which denotes the indirect perpetration of a crime through an organized apparatus of power, or a "machine". In order to establish *organisationsherrschaft*, it must be proven that the defendant's orders are carried out as a matter of course by the individuals who are "cogs" in the criminal organisation or "machine" (*ibid*. 7). In this case, the individuals themselves are not found culpable but the orchestrators are. If this doctrine was applied to AWS, it would mean that the commanding officer could be found culpable for the actions of AWS.

Ohlin argues that the doctrine of JCE could be applied to a machine, the only change that would need to be made would be to take the metaphoric language as literal, in other words, to view the weapons system as a the "machine" organisation working under the control of the military commander (*ibid*. 8). Critics of this idea could object that this is improper because AWS have enough autonomy that they could not literally be viewed as an automated machine, but Ohlin defends his proposal by claiming that this is precisely how AWS are in fact similar to human "cogs", and indeed the mechanisms of *organisationsherrschaft* requires autonomous agents that act as a part of a "machine" (*ibid*.). Ohlin does concede that the limiting principle here is that the instrument has to be non-culpable, in other words, functionally equivalent to a real instrument (2016: 10). This means that the individual "cogs", or the AWS, will always have limited or no culpability, which is something that we do not want. The reason that this is undesirable is because it denies that the individual perpetrators of the act any blame and allocates responsibility solely to the organisers, which challenges our ordinary moral intuitions. This proved to be an issue at the outset of the Eichmann case (*ibid*. 10).

Otto Eichmann was one of the major organisers of the Holocaust. In this instance, many people acting on his commands worked together to produce the horrific outcome of the

Holocaust and the war crimes in question. During Eichmann's trial, applying the rule of *organisationsherrschaft*, and convicting him on this basis, would have had the undesirable consequence of denying the culpability of those who executed his orders, and many believed that would have been too high a jurisprudential price (*ibid*. 11). Thus, Eichmann's prosecutors modified the rule of *organisationsherrschaft* so that its application did not necessarily exclude the culpability of the "cogs" in the machine, such that both Eichmann and those would carried out his orders could be held accountable. The result, for the purposes of our discussion, is that a blameworthy commanding officer could be found culpable according to the doctrines of the JCE and *organisationsherrschaft* and this would not affect the culpability of AWS.

Using both the examples of the Nuremberg trials and the Eichmann case, Ohlin outlines existing legal precedent where the commander could be held liable whether the AWS is a responsible moral agent or not, and whether the AWS satisfies the Combatant's Stance or not (*ibid*. 19). The limitations of these mechanisms are that they only apply when the commanding officer is acting *mala fide* and is culpable. While it is possible that this could occur, it is likely that the far more common instance is one where the commander who acts in a *bona fide* manner deploys the AWS for military operations, and the AWS violates one of the core prohibitions of IHL (distinction, necessity or proportionality) (*ibid*. 20). In these cases, the existing legal mechanisms fail to guide the establishment of responsibility, as will be shown. If the commanding officer were to be aware of the probability of this outcome, and chooses to deploy anyway, he could be considered to be reckless (*ibid*.). Recklessness is a concept that is unfortunately absent in international criminal law, and there is no suitable existing method to prosecute a reckless commanding officer for the commission of a war crime perpetrated by AWS. International criminal law has what Ohlin refers to as "a serious blind spot" where crimes of recklessness are concerned, or what civil law calls *dolus eventualis* (*ibid*.).[140]

Legal scholars overwhelmingly agree that recklessness is a less culpable mental state than *mens rea*, or malicious intent, since the act was not purposeful. Some international

[140] The liability for the risk of a future event is recognised and a concrete decision to move forward (even in light of the potential for a negative outcome) was made. If saying the Reasonable Man would have foreseen this probability of this outcome occurring, then the defendant is guilty *dolus eventualis*.

humanitarian lawyers believe that certain war crimes are governed by the less culpable mental state such as recklessness or *dolus eventualis*, on the basis that they often lack the necessary intention to do harm (*ibid*. 21).[141] While there are at least some interpretations of command responsibility that allow for a military commander to be prosecuted for crimes committed by subordinates, they are limited to circumstances where the military commander is deemed reckless due to him failing to fulfil his responsibility to investigate potential crimes and to hand out punishment (*ibid*.). In this instance, the failure to supervise or punish troops under one's command is considered to be an instance of recklessness. This notion of recklessness is catered for in the JCE, since the defendant can be convicted for crimes performed by the members of the JCE, including crimes that are not a part of the original criminal design, because the designers would have had to foresee the potential of members straying from the original plan (*ibid*. 22). The problem in convicting a reckless commander under the JCE is that the doctrine does not recognise lower-level mental states like recklessness in the grading of the offense, since the definition of the offense reflects the culpable mental state.[142] Ohlin argues that this seems wrong, as the legislative grading of offenses[143] based on culpability is not only pragmatic but also morally imperative (*ibid*.). Ohlin argues that we all have a reasonable expectation that the criminal categories that are attached to conduct will adequately reflect the state of the criminality of the act and not unreasonably equate significant differences in culpability (*ibid*. 23). Some believe that this can be accounted for by judges in the sentencing portion of the trail by having the judge impose a less severe sentence based on a lower-level of culpability (*ibid*.). This misses the point of requiring legislative and pre-existing definitions of offenses, which serve to guide judges in sentencing. Thus, there are good grounds to examine issues of recklessness and AWS more closely.

The principal international crimes are crimes of aggression, genocide, crimes against humanity, and war crimes, and, unfortunately, international criminal law does not allow for specific offences that differentiate between intentional crimes in these categories and reckless

[141] For example, failing to adequately distinguish targets in the heat of the moment is unlikely to be considered intentional. It is still a war crime, though many would agree that it is a less culpable act than intentionally targeting civilians.
[142] For example, failure to discriminate between combatants and civilians presupposes that the civilians were killed intentionally, rather than accidently.
[143] For example, the distinction between murder (an act of intentional killing) and culpable homicide (an act of negligent killing).

ones (*ibid*. 24). This means that a negligent or reckless commanding officer may be treated as if they were intentionally malicious. Ohlin believes what is needed is an entirely new criminal offense that distinguishes between an intentional war criminal and a recklessness war criminal, or distinguishes between someone whose conduct carries the risk of being a crime against humanity, from someone who intentionally carried out such a crime (*ibid*.). After all, while all military operations run some risk of running awry, it only seems fair to punish those who execute a military operation despite it having a significantly high risk of derailing. In other words, if a specific AWS is deemed unsuitable for the operation, or has some errors that make it less reliable than normal, and the commander deployed the AWS regardless, they ought to bear some degree of responsibility. However, if the operation is one that usually is suitable for AWS, and the AWS has a high degree of reliability, the commanding officer is justified in deploying it and should bear no criminal liability for AWS deviance. Hence Ohlin believes that it seems counterintuitive that we thoroughly police levels of culpability by distinguishing between principal perpetrator, co-perpetrators and accessories to actions, but we ignore the similarly important distinction between intentional and reckless conduct in international criminal law (*ibid*. 25). If it is indeed likely that most instances of war crimes committed by AWS will be due to reckless conduct rather than intentional behaviour[144], there will be problems for the prosecution of a commander who has deployed an AWS that commits a war crime (*ibid*.). It is not appropriate to convict a reckless military commander for the exact same crime as the war criminal who intentionally targets civilians or intentionally executes a fallen soldier who is *hors de combat* or a prisoner of war (*ibid*.). But, while this hypothetical commander is less culpable than one with intent to commit a war crime, it would be wrong to say he or she is not culpable at all, since it is his or her responsibility to exert a certain degree of control over those under his or her command. Such a commander ought to be prosecuted of *something*; he or she has failed his or her responsibility to control the operation (*ibid*. 26). Ohlin proposes that one possibility to deal with this issue is to re-purpose the distinction between principal perpetrators and accomplices, such that a reckless commander would be viewed as an accomplice to an AWS (*ibid*.). Ohlin recognises that a potential downside to this is that it would require the AWS to be viewed as the principal perpetrator and thus as a morally responsible agent (*ibid*.). The problem with this is that it implies that the AWS is essentially a

[144] This seems likely because of the possibility of AWS at times acting in ways that the commanding officer cannot predict, thus they may not be able to control AWS all of the time.

person, which is a hard pill for some to swallow. It would cause less contention in the academic community to find a solution that does not require this.

A better solution that Ohlin proposes, albeit a more time-consuming one, would be to codify an entirely new criminal offense, say for recklessly perpetrating an international crime, such that the military commander who recklessly deploys an AWS that violates IHL would be guilty of a specific AWS-related criminal offence (*ibid*. 26-7). Ohlin believes that the downside here is that such legislative solutions are unlikely in a system controlled by the Assembly of State Parties at the International Criminal Court, an institution unlikely to consider substantial revisions to the Rome Statute (*ibid*. 27).[145] Nevertheless, despite the resistance shown to amending the Rome Statute, some legislative amendment is surely required if and when AWS (and their associated war crimes) become a reality (*ibid*.).

The suggestions presented by Ohlin represent relatively modest changes to existing structures that could be implemented relatively quickly, and, as such, they are worthy of consideration. Ohlin makes a strong case for asserting that the basic structure of international criminal law is already well suited to prosecuting military commanders for the deployment of AWS that commit war crimes, by the literal interpretation of the metaphor of the organised "machine" and by adopting the combatant's stance. The only area lacking is the recognition of crimes of recklessness by international criminal law (*ibid*. 28). Nonetheless, Sparrow's assertion that we can never legally assign blame to the commanding officer is refuted. There will, however, be instances where the commanding officer is neither culpable nor reckless, and it would not be possible or fair to hold him or her liable. In those situations, where neither the programmers nor the commanding officer can be blamed, the responsibility for the commission of a war crime must lie somewhere else.

[145] The Rome Statute is what established the International Criminal Court, and as a result they have a bias against amending it.

5.3. The Machine

If we cannot assign responsibility for a war crime to the programmers or to the commanding officer, then the last potential candidate for taking responsibility is the machine itself. Sparrow (2007) finds the idea that we ought to hold the machine accountable absurd. While he acknowledges that we could say that it is causally responsible for a war crime, he believes it is a step too far to hold it morally responsible and liable as well, since that would imply that it is an artificial agent (Sparrow, 2007: 71). It is a conceptual struggle to imagine exactly how we might hold the machine liable in a meaningful manner, thus most would accept Sparrow's sentiment here. As Asaro (2011: 181) explains, the issue is that criminal actions necessarily require a moral agent to perform them and it is not clear that a machine qualifies as a moral agent. After all, when a person is considered responsible, it means that we necessarily find them to be the best available candidate for reward or praise and consequentially of punishment and blame, and Sparrow asserts that there is no reasonably manner for us to reward or punish a machine. Asaro (180) asserts that it is unlikely that we will resolve the issue of machine moral agency anytime soon, so establishing mechanisms for machine culpability would initially need to be independent from the issue of personhood.

For some, the inconceivability of meaningfully chastising an AWS for its behaviour, and thus the implausibility of holding it morally liable, provides enough grounds for banning AWS outright. However, a counter-argument is that if the machine is autonomous and intelligent, and has the required internal motivations and desires, it would not be unreasonable to hold it liable. The claim is that if machines have intellectual capacities similar to those of a human agent, it is not unreasonable to treat them like artificial agents, and punish them the way we would punish human soldiers who violate the Rules of Engagement (Sparrow, 2007: 72). Sparrow counters this claim by saying that this would only satisfy our psychological need for justice but would not be the correct locus of responsibility.

Sparrow asserts that if one were to hold AWS responsible for its behaviour and punish it accordingly, that the punishment it receives must be meaningful (*ibid.*). Sparrow believes that in order for the punishment to be meaningful, the machine should evoke feelings similar

to those induced by other human agents – sentiments like compassion, understanding, and empathy (*ibid*.). This means that it ought to "suffer" in a capacity that motivates the same set of responses in a human, to the extent that if we discovered it were innocent, we would feel remorse for the injustice it suffered, and we would owe it some recompense. Essentially, they would be full moral persons, and we would treat them as such. But Sparrow argues that if this were the case, it would undermine one of the goals of creating an AWS, namely to limit risk to our soldiers (*ibid*. 73). One of the reasons for the movement towards the automation of weapons is because taking humans out of the loop makes them safer, and if AWS were artificial agents, we would just be putting "people" back in danger. Sparrow believes there are two alternatives here: either machines cannot be held liable because they are not moral agents, or if they can be held liable, and they are such moral agents, it would undermine the initial goal of their development (*ibid*.). Either way, Sparrow believes that the issue of responsibility is unresolvable, and therefore we ought to not field AWS. I disagree with Sparrow that the creation of an AWS with this capacity undermines the goal of their development; most AWS will have weapons hardware in the field but the "brain" that makes the decisions (and thus the part that could have some level of "intentionality") will be located somewhere outside the combat zone. Furthermore, the AWS that do carry their "brain" with them will be more durable than flesh and bone humans are, and in fact it even more will likely be non-physical software. All of these points undermine Sparrow's claim that the use of AWS with intentionality undermines the goal of making moral agents safer. Additionally, if AWS are more precise and perform better than human agents, like Arkin believes, a lack of direct human-like liability, or the implausibility of meaningfully holding a machine accountable, is not enough to motivate an outright ban. In this case, the use of AWS will reduce the casualties of war, and there are methods of assigning responsibility to other parties that will be sufficient to ensure their lawful and ethical use.

Sparrow attempts to further his claim that the lack of responsibility justifies an outright ban by comparing AWS with child soldiers. He writes that one of the reasons we find the use of child soldiers unethical is because they have limited liability and cannot be held accountable for their actions (*ibid*.). They lack full moral autonomy, but they are still considered to be autonomous agents. I disagree with Sparrow's comparison. A child's lack of full moral responsibility is not the only reason we find the use of child soldiers unethical, and it is not

even the primary one. The use of child soldiers is unethical because children are considered persons with special protections and rights by the United Nations[146] and the conscription of a child under the age of 18 constitutes a war crime. Children cannot meaningfully give consent and thus cannot be legally conscripted into the military. Further, Sparrow's assessment that children cannot be held liable is untrue; while there is a presumption that children under a certain age lack full intentionality and thus liability, there are exceptions to this presumption and there are many cases where children have been held criminally liable for their actions.[147] His comparison is inappropriate. Over and above this objection, based on arguments presented by Dennett and Allan and Wallach, I want to argue that there are ways in which we can hold machines meaningfully responsible, thus refuting Sparrow's position.

Daniel Dennett (1997: 353) takes the example of Deep Blue to show how we might assign praise and blame, and consequentially, responsibility to machines. Deep Blue is an expert system[148] designated as a chess-playing computer engineered by IBM. It is the first digital computer to attain victory over a world champion, Gary Kasparov, in both a chess game and a tournament under league conditions. The significance of Deep Blue's victory is that it refuted claims made by critics of strong AI that a machine, being limited to syntactic rule-based intelligence, would not be capable of making the complex probabilistic decisions necessary to outwit a human agent. In his paper (*ibid.*), Dennett asks who we might praise for this victory in order to establish whether a machine could be considered to be an intentional agent and, by implication, whether it is possible to hold it liable. The programmers of Deep Blue did not beat Kasparov, nor did they know what the winning sequence of moves that won the game would be; only Deep Blue knew its strategy and adapted its tactics to defeat Kasparov (*ibid.*). Dennett concludes that while we may congratulate the programmers, in the same sense that we would congratulate Kasparov's teachers and parents, ultimately Deep Blue is the best candidate for us to credit the victory to (*ibid.*). Dennett continues that this implies its behaviour is predictable and explicable if we attribute cognitive states[149] and motivational states[150] to it, which is what

[146] The rights of children are specifically codified in The Convention on the Rights of the Child. The Additional Protocols to the Geneva Convention criminalises the conscription of child soldiers.
[147] One example is Lionel Tate, who was 13 when he was convicted for murder and received a life sentence without the possibility for parole (Canedy, 2001).
[148] Similar to the expert systems envisioned by Turing discussed in Chapter 2.
[149] Equivalent to beliefs.
[150] Comparable to desires.

we do when we attribute Deep Blue with the faculty of reason needed to determine what action to pursue in order to attain those desires (*ibid*.). This, he argues, makes Deep Blue an intentional[151] system (*ibid*.). Dennett's logic echoes Ohlin's assertion that we ought to view AWS with the Combatant's Stance. According to Dennett, we can say that Deep Blue had beliefs and desires about its activities on the chessboard, and in light of this, Deep Blue is the best candidate for being the responsible agent who defeated Kasparov (*ibid*.). Similarly then, Deep Blue should then be the one who we ought to blame for a loss. In the context of AWS, if the AWS had desires such as specific military goals, and beliefs about potential targets and strategies, and could act in accordance with its beliefs to attain the desired outcome, the AWS has intentionality according to Dennett's argument. This would make the AWS the logical candidate for praise and blame. I agree with Dennett's position that AI acting with the relevant beliefs and desires is intentional, in the philosophical sense.[152] However, Dennett continues that this level of intentionality is not enough for full moral agency or responsibility. Many scholars like Dennett believe some further understanding of the moral value of the action is required, however, as I discussed in Chapter 3, in criminal law intentionality has little to do with having beliefs or desires about the world, but rather refers to acting with purpose. The argument that I make is that this lower level of intentionality (of having certain beliefs and desires) is sufficient for legal purposes. Regardless, assigning responsibility to AWS need not be limited to AI with either philosophical or legal intentionality; there are other legal practices that can be used to punish AWS, and Asaro discusses a few of them.

Before examining the existing legal practices that can be adapted to hold AWS culpable, Asaro (2011) first examines why we punish criminals. He begins by examining the reasons for punishing people, identifying three principal motivations: retribution, reformation and deterrence (*ibid*. 181). It is certainly possible to punish a machine out of a sense of retribution, but as Sparrow mentioned, this merely satisfies our own psychological need for revenge. Asaro continues that it seems strange to say we should punish AWS to reform them, since only a moral agent can be reformed, and machines are not moral agents (*ibid*.). Asaro similarly argues that punishment as a mechanism for deterrence is only applicable when the recipient of

[151] This concept was discussed in the Chapter 2.
[152] Dennett's position is that we are treating AWS as if it had beliefs and desires, rather than stating AWS actually has such beliefs and desires. This logic is similar to Ohlin's argument for the Combatant's Stance.

punishment is a moral agent that recognises other moral agents (*ibid*.). In other words, a moral agent is only deterred from acting in a criminal manner if they see other agents being punished for that behaviour and are motivated to avoid similar consequences. Asaro believes that at the very least, there should be a measure to destroy AWS that do harm, but just as with animals the purpose will not be to achieve retribution, reformation or deterrence, but more as a preventative measure to stop further harm (*ibid*. 182). Thus destroying rogue AWS will not conform to the traditional reasons for punishment. Nevertheless, Asaro is correct in his assertion that rogue AWS need to be decommissioned, since they are no longer reliable and predictable and thus can no longer be used in compliance with IHL. There are, however, methods of punishing AWS that do conform to the traditional reasons, which is significant because it gives the punishment meaning over and above the mere satiation of our desire for revenge, but I will return to this shortly.

The problem of assigning responsibly to non-human entities first presented with the introduction of corporations in the 17th century. The Lord Chancellor of England at the time said that these entities have "no soul to damn and no body to kick", thus they would be incapable of being held accountable for their actions (*ibid*.). In this case the issue was addressed with the concept of monetary punishment; corporations were fined as a form of punishment. In cases where the offence was severe enough, corporations could be split apart into smaller entities with less power, or dissolved altogether. While the corporation cannot be imprisoned, the individual actors within it can be prosecuted (*ibid*.). A possible solution the responsibility dilemma with regard to AWS is to treat them like corporations, that is to say, as nonhuman entities with legal capacity. AWS could also be compartmentalised, much like a corporation is split, or deactivated all together. Certain military personnel would serve as the "board of directors" so to speak, and they would be held liable for crimes committed by the AWS, but less so than would be the case where they were personally culpable.

In addition to treating AWS as legal entities for the purposes of punishment, there are ways to punish the machine itself in a manner that conform to the traditional justifications for punishment. Drawing on an argument presented by Johnson, I believe punishment as a mechanism for achieving reformation and deterrence could be possible. She believes that AWS

could be designed in a way that they are responsive to the mechanisms of reward and punishment, comparable to reinforcement learning (Johnson, 2014: 3). Her assertion is not that machines or systems would truly *be* responsible, but instead that humans will be disposed towards considering and interacting with them *as if* they were responsible (*ibid*.). In this way, punishment would form part of their moral education and learning, such that any instances of punishment teach the machine to not repeat that action. Indeed, punishment and reward is a mechanism of machine learning; rewards reinforces that an action is desirable and should be performed while punishment reinforces that an action should not be performed. This could also deter machines from behaving in a way that results in punishment; as AWS become more sophisticated, theoretically they can learn from the punishment inflicted upon other systems, and so be deterred from making the mistakes that the punished AWS made. I believe that we can punish machines in a manner that is consistent with traditional theories regarding the purpose of punishment, which legitimises punishing the machine and punishes it in a manner that is meaningful to us. [153] In other words, it may be possible to punish machines as a mechanism for retribution in order to satisfy our own need for revenge, or as a means of reformation or deterrence, as proposed by Johnson. Thus, there are several scenarios discussed above where it is possible to effectively and meaningfully hold AWS culpable, which negates Sparrow's argument. Thus if AWS can be held responsible in a manner that is meaningful to us, there is no substance to the objection that there are no suitable candidates for responsibility for the actions of AWS, thus eradicating a significant obstacle to regulating their use.

5.4 Summary

The controversial nature of machine liability is a recurring criticism against the use of AWS and something that needs to be addressed. There are many authors that consider various loci of responsibility and others that find the discourse on the matter lacking. It is important to remember that maintaining direct causal responsibility between an individual and a criminal act is not always necessary for allocating responsibility for that act, especially under international law. Robert Sparrow considered three obvious loci for responsibility, the programmers, the commanding officer, or the human with "meaningful control", and the

[153] For example, AWS could be programmed with a feedback mechanism that cause them to avoid behaviour has previously resulted in "punishment" for them or for other AWS.

machine itself, all of which he rejects as being appropriate. I only agree that it would be inappropriate to always identify any one of these as the responsible candidate, but I have argued that there will be various instances where one of the candidates is the most appropriate and they should then be held liable.

Programmers could be held liable in accordance with existing product liability law, as Asaro suggested. This would be effective immediately and would not require the drafting of new legislation. As mentioned, the only issue that needs to be addressed is the development of "industry standards" for the manufacture of AWS. With regard to the commanding officer, Asaro argued that AWS could legally be treated like domesticated animals, where the commander would be liable for damages AWS inflicted and would be criminally culpable if their command was negligent or reckless. It will be the operator's responsibility to properly train these systems and to make suitable judgement calls about whether it is appropriate to deploy them (Allen and Wallach, 2012: 127). Further, Ohlin made a compelling argument that the international criminal law mechanisms established during the Nuremberg trials and in the Eichmann case could be employed to prosecute commanders who acted criminally intentionally. Ohlin points out that crimes of recklessness, which is a less "intentional" culpable state, is an area where international criminal law is lacking, and that the best solution would be to codify a new criminal offence for recklessness that facilitated the perpetration of a war crime. These issues need to be addressed in the codification of arms control regulation. Furthermore, I argued that it is possible to hold AWS criminally liable, in a manner that is consistent with traditional legal philosophy. This makes the punishment of AWS a legitimate and viable option. Therefore, there are no grounds to ban AWS due to any "responsibility gap".

6. Conclusion

Given the trend in weapons automation and the strides in the development of AI, the development of fully autonomous weapons systems is inevitable despite apprehensions held by opponents. This is a polarising technology; some are stalwartly opposed to them, claiming that their use is immoral and unlawful, while others believe AWS could offer great benefits. But ultimately, we cannot go down a pros and cons list and subsequently ban or employ AWS categorically. It is unlikely that these weapons will be banned, and even if the United Nations or a majority of countries adopt this policy, enforcing it will be largely impossible (as is the case with missiles and landmines). Instead of focusing on the costs versus benefits of AWS, the discussion needs to focus on regulation to ensure that the use of AWS complies with IHL and that all associated issues, like responsibility, are clearly addressed. And given how long adequate regulation takes, the discussion needs to shift there as soon as possible, so that there are no ethical or legal lapses that allow militaries to exploit the ambiguities relating to proportionality, responsibility and recklessness in order to run risky and harmful operations with less risk of criminal culpability. To reiterate, it is no longer useful to discuss an outright ban on the use of AWS; rather academics and policy-makers ought to focus on regulation.

An integral part of understanding the nature of AWS and the debate surrounding them is understanding artificial intelligence, since AI is central to the automation of tasks. Thus, in order to make the nature of the debate surrounding AWS more understandable, I first gave a reductionist overview of AI theory, which is essentially the theory of intelligence, and discussed the two most basic perspectives. First, I summarised Turing's functionalist position that the mind is a type of machine, and without access to the internal mental states of others, we deduce the existence of their consciousness and intelligence on the basis of their behaviour. Next, I showed that this behavioural equivalence is not a sufficient ground on which to attribute intelligence for many AI theorists and that many believe some other property is required to qualify for human-like intelligence. Most notably, Searle argued that humans have an extra property, namely intentionality, which is the cradle of intelligence. However, I noted that intentionality is an internal mental property and Turing correctly noted that we cannot know of

its existence in others, thus it is an unusable measure for intelligence. Instead I advocated for Ohlin's position, that in the absence of knowledge of the internal mental properties of others (human or machine), behavioural equivalence is all we have to go on. Concerning AWS, this would require us to adopt the Combatant's Stance when interacting with them. That is, if an enemy AWS behaved in a manner that we could not tell whether the enemy was human or not, we would have to assume it is human in order to interact with it in a manner that make sense to us. I took Ohlin's position further, to argue that if an AWS is able to make the same decisions about the application of lethal force as a human soldier in compliance with IHL, then we would have to afford it the same decision making responsibilities as we afford to a human soldier, for the sake of logical consistency. I supported this position by referring to Frowe's distinction between the right to inflict harm and the liability to bear that harm to show that the decision to apply lethal force can be separated from the moral liability of the harm itself. Specifically, I used Frowe's work to argue that the internal mental properties of the subject, or Searlean intentionality, are irrelevant when determining whether there are grounds for lethal action on the part of the AWS, and these properties only come into play in matters of responsibility. Thus in Chapter 2 I provided an argument whereby AWS could be granted the same authority to make lethal decisions as humans, provided they are able to make these decisions as well as (or as consistent with IHL) as human soldiers, without requiring that the AWS in question has intentionality or is capable of being a responsible agent.

After furnishing a basic understanding of AI, I went on to discuss the various definitions of AWS that a few salient organisations[154] hold, in order to fashion a definition that can be broadly accepted and applied. The reason establishing a firmer definition is important is because it affects where the burden of responsibility lies. A weapons system that is only partially autonomous is legally similar to a weapons system that has no autonomy, since a human operator must approve the decision to employ lethal force. When examining the various definitions held by the relevant organisations, it is apparent that an AWS is defined by its ability to fulfil two critical functions: the ability and freedom to identify and select targets, and the ability and authority to use lethal force. This definition is broad enough to include a wide variety of fully autonomous weapons systems while excluding ones that are only partially

[154] The specific organisations I mentioned where the US DoD, the Human Rights Watch and the International Committee of the Red Cross.

autonomous. The exclusion of partially autonomous weapons is significant because it implies a difference in the level of responsibility that can be accorded; the human operator is the sole locus of responsibility in the case of a partially autonomous weapon. And these types of fully autonomous weapons systems are the ones that present the most ethical and legal issues, and are thus the ones I focused on.

One of the concerns I identified that AWS raise is that it is difficult to maintain the causal chain of responsibility. Currently, the solution that the US DoD has employed to deal with this issue is to require "meaningful human control" over the use of AWS. That requires a human supervisor who acts as a watchdog and is essentially responsible for failing to perform their duty as a supervisor. There is an implicit difference in the level of culpability between a human operating a weapons system and one who is merely supervising it; the former is more culpable for violations of IHL than the latter. This is essentially a form of regulation. However, I found this method of regulation to be inadequate, for two reasons. Firstly, what exactly constitutes "meaningful human control" is left vague and is there is no international legal prescription requiring human supervision. If this is to be a standard of arms control, it needs to be clearly defined. Secondly, the US DoD imposed this regulation on itself, which means it may revoke it at any time. I argued that any methods of regulation need to be required by IHL in order to be effective. Given that "meaningful human control" could plausibly be used as method of effective arms regulation (provided it is internationally adopted), I gave an account of what could constitute this kind of meaningful control. Specifically, I recounted Asaro's argument that "meaningful human control" implies that the system performs sufficiently predictably and reliably and that the human has the ability to intervene if need be.[155] Asaro also asserts that the operator can only be held accountable if he or she were afforded meaningful control, and were not reduced to fulfilling a role as an approval mechanism in an ineffectual attempt to comply with IHL. I proposed that Asaro's explanation of "meaningful human control", or something similar, should be a starting point for formal arms control measures. This is an easy way of ensuring compliance with the standards of human decision making outlined in IHL; not only does it ensure that someone is present to override the AWS if need be, it also provides a candidate for responsibility for the actions of the AWS. Furthermore, I

[155] Human intervention could be subtle, such as manually switching targets, or a more overt takeover of all the AWS' systems.

have also discussed the ICRC's recommendations that can be implemented unilaterally by governments in lieu of international consensus, which have a similar regulatory effect to Asaro's proposal. While it is unlikely that governments will take such actions without international pressure, measures like limiting the functions or environmental range are good examples of arms control measures that may be codified in IHL. Furthermore, these measures maintain the autonomy of the weapons system, thus are applicable to the types of systems discussed throughout this paper.

I also discussed some of the reasons why we would want to use AWS, or rather, some of the benefits that AWS offer. In short, AWS are theoretically capable of responding faster than humans, which presumably translates more effective acts of defence. Additionally, AWS engineered to lack human weaknesses, both psychological and physical, would not be inclined to act out of fear or anger or similar emotions, which will no doubt mean stricter compliance with the ROE and IHL. Both of these factors are likely to result in saving more lives and reducing collateral damage, which I argued are very good reasons to employ AWS, subject to AWS being utilised in a manner that complies with IHL. The regulation of AWS must keep these capabilities in mind, in order to ensure that whatever limitations are placed on the use of AWS do not erode these benefits. However, I also noted that if AWS are to be used then developers of AWS must overcome the principal objections to the use of AWS. The two principal objections I identified where firstly that the use of AWS is *de facto* incapable of complying with IHL and secondly that there are no suitable candidates for responsibility in the even that an AWS violates one of the principles of IHL.

As for the first objection, I reviewed the circumstances under which IHL permits the use of lethal force, examining the two guiding principles for the use of lethal force; specifically the principle of discrimination and the principle of proportionality. I argued that there is no convincing argument that proves that AWS are inherently incapable of discriminating between combatants and civilians based on direct participation in hostilities, as this determination is made on the observation of the acts of an individual. Thus any AWS capable of determining that an individual is performing an act that meets the required threshold of causing harm in one causal step and is performing the said act with *belligerent nexus*, as discussed in Chapter 4,

legally has the ability to decide whether or not to employ lethal force. I also argued that given the groundwork I laid in Chapter 2, that this is independent from whether or not the AWS is capable of bearing criminal responsibility. I also argued that in the event that the AWS is unable to accurately make this distinction, the human operator with the burden of "meaningful control" can make this determination on behalf of the AWS and prompt the act. I further wrote that limiting AWS autonomy is an acceptable trade-off to ensure compliance with IHL.

The second guiding principle is that of proportionality, which is where I found IHL to be lacking; what constitutes a proportional response is left too vague. I have argued that this principle needs to be better articulated, even if only to better guide human soldiers. The examination I presented shows that the principle overlaps in some ways to the lack of discrimination, but this cannot completely define proportionality. Furthermore, I argued that proportionality is not restricted to obviously excessive methods that are already banned, like chemical and biological weapons. The best approximation I could find, and one argued for by the Human Rights Watch, was in the 1990 Basic Principles on the use of Force and Firearms, that is, force must be necessary, must constitute last resort, and should not be disproportionate to any advantage gained. I argued that it comes down to a utility calculation, and that if human soldiers are trained to make these kinds of decisions, there is no reason an AWS would not be able to do so as well. Again, in the intermediate stage, the human operator should ensure that the use of force is proportional. To reiterate, my examination of IHL in light of the use of AWS has shown that it is possible for an AWS to be used in compliance with IHL, and the human operator, at least at the initial phases of deployment, will monitor the AWS to ensure compliance. The only two issues that I have found are in the vagueness of the principle of proportionality and in the conceptualization of "meaningful human control". Both these concepts need to be more expressly articulated and formalised in legislation. With regards to the permissible use of lethal force, arms control needs to require either the AWS of the human with meaningful control adhere to the principle of distinction and proportionality. Part of this process will mean defining proportionality more precisely.

The second challenge to overcome is the issue of who to hold responsible when an AWS contravenes IHL. Opponents to the use of AWS argue that there is a responsibility gap

between the agent who performs the action and the one who must bear responsibility for it. I have examined Sparrow's argument on the various loci of responsibility for a war crime committed by an AWS. I agree with Sparrow that negligent or reckless programmers can be held responsible, and supplemented my position with Asaro's argument that existing mechanisms of product liability law could be used to hold them responsible. However, product liability law hinges on the existence of industry standards. As such, there is a need to develop industry standards to give the engineers of AWS a measure to conform to, such that there is a measure for what constitutes negligence or recklessness on the part of the manufacturer. Secondly, I rejected Sparrow's assertion that the commanding officer cannot be held responsible, since currently the human operator is the one who is responsible for monitoring the AWS, and this seems to be the norm that is building. If we take existing legal precedents established during the Nuremberg trials, the Eichmann case, and the establishing of the Joint Criminal Enterprise, we most certainly can hold a commanding officer liable if his behaviour were culpable without having to make drastic changes to the existing legal system. The issue I noted is that most commanding officers will not be intentional war criminals; rather, they will be negligent or reckless, and this is an area that Ohlin points out that IHL is lacking. There is a gap between what happens in practice and the law, which means that reckless commanding officers will be either over- or under-punished. The best solution to ensure this legal gap is covered is to criminalise new offences regarding the irresponsible supervision of AWS and reckless commanding officers. Lastly, I also rejected Sparrow's assertion that it is absurd to hold the machine liable on the basis of the questionable nature of machine intentionality. If it were an intentional agent, we would definitely treat it as a full moral agent, and if it were not, we could still hypothetically treat is as if were liable, using praise and blame to reinforce moral education. If one looks at Dennett's claim that Deep Blue is the best candidate for praise for its chess prowess, then the machine could similarly be the best candidate for praise or blame for operational outcomes. I have also offered Johnson's proposal that AWS could be engineered in a manner that they are responsive to praise and blame, so that we can "punish" them so as to ensure reformation and deterrence. I have also explored other various legal avenues that bridge the so-called responsibility gap. One such solution is by modifying the law of State Responsibility, such that the State could be held responsible for failing to police the use of AWS effectively by the international community. Another possibility is to legally regarding AWS in the same manner as dangerous pets, such that the supervisor is expected to have a reasonable level of control and responsibility over the AWS without assuming they ought to

have full control at all times. To reiterate, there are mechanisms that can be used to hold the programmers, the commanding officer, or the machine itself responsible, depending on the circumstances; the only major problem lies with the lack of legislation dealing with negligent or reckless commanding officers. In respect to responsibility, arms control regulation needs to ensure that the mechanisms for holding an agent – whether it is the manufacturer, the commanding officer, or the AWS itself – responsible are clearly defined, as well as in which circumstances it is appropriate to hold which agent liable.

In conclusion, I have shown that the principal objections cited by opponents are insufficient grounds for an outright ban and that these objections can be overcome with relative ease and only a few adjustments to the existing legal system. I conclude that AWS are not inherently incapable of conforming to IHL, which requires discriminating between combatants and civilians, and a proportionate application of force. I have argued that any AWS that is able to identify a target as directly participating in the hostilities (on the basis of threshold of harm, direct causation, and belligerent nexus) as well as a human soldier should be considered capable of adhering to the principle of distinction. Furthermore, I offered a potential criterion for which to measure proportionate response by, not only to give the developers of AWS a starting point for engineering purposes, but potentially also to reform IHL and give human soldiers a firmer idea of what proportionate force is. If an AWS is able to determine that force is necessary, that it is a last resort, and that in is proportionate according to a utility calculus, to a similar degree of accuracy as a human, then that AWS is capable of adhering to the principle of proportionality. I also conclude that there are instances where the manufacturer, the commanding officer or the AWS itself may be held meaningfully liable, on the grounds of existing product liability law, the laws regulating corporations and the ownership of dangerous animals, the JCE and the Nuremberg and Eichmann trails, and mechanisms of machine learning that make punishment meaningful to the AWS. All of the suggestions I offered could be incorporated into arms control regulation, such that it is clear what type of AWS are afforded the decision to employ lethal force, and all matters of responsibility are clearly defined. This is a sufficient starting point to dealing with the issue of arms control over AWS. Areas of future research would include providing a clear taxonomy of industry standards for AWS manufacture, and examining the use of AWS in broader circumstances outside of armed conflict (for example, in general peacekeeping capacity or in the private sector).

Bibliography

Allen, C. & Wallach, W. 2013. *Framing Robots Arms Control*. [Online] Available: https://www.law.upenn.edu/live/files/3395-allen-c-wallach-w-framing-robot-arms-control-2013 [15 July 2016].

Anderson, K. & Waxman, M. 2013. *Law and Ethics for Autonomous Weapons Systems: Why a Ban Won't Work and How the Laws of War Can*. [Online] Available: https://www.law.upenn.edu/live/files/3392-anderson-k-waxman-m-law-and-ethics-for-autonomous [3 June 2016].

Arkin, R. 2013. *Lethal Autonomous Systems and the Plight of the Non-Combatant*. [Online. Available: https://www.law.upenn.edu/live/files/3880-arkinlethal-autonomous-systems-and-the-plight-of [18 May 2016].

Asaro, P. 2012. On Banning Autonomous Weapons Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision Making. *International Red Cross*, 94 (886): 687-709.

Asaro, P. 2016. Jus Nascendi, Robotic Weapons and the Martens Clause (in press).

Asaro, P. 2011. A Body to Kick, But Still No Soul to Damn: Legal Perspectives on Robotics. In Patrick Lin, Keith Abney and George Bekey (eds.) *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA. MIT Press. 196-186.

Aspen Institute: What is Agent Orange? [Online] Available: https://www.aspeninstitute.org/programs/agent-orange-in-vietnam-program/what-is-agent-orange/ [18 October 2016].

Barker-Plumber, D. 2016. *Turing Machines.* Stanford Encyclopaedia of Philosophy. [Online] Available: http://plato.stanford.edu/entries/turing-machine/ [8 October 2016].

Blackburn, S. 2009. Can Machines Think?, in, Blackburn, S. (eds.). *The Big Questions*. London: Quercus Publishing Plc. 85-93.

Boden, M. 1990. Introduction, in, Boden, M. (eds.). *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press. 1-21.

Boden, M. 1990. Escaping the Chinese Room., in, Boden, M. (eds.). *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press. 89-104.

Cummings, M.L. 2014.*The Human Role in Autonomous Weapon Design and Deployment.* University of Pennsylvania Law School, [Online] p. 1-14. Available: https://www.law.upenn.edu/live/files/3884-cummings-the-human-role-in-autonomous-weapons [16 March 2016].

Dennett, D.C. 1997. When HAL Kills, Who's to Blame? Computer Ethics, in: Stork, D. (eds.). *Hal's Legacy: 2001's Computer as Dream and Reality*. Cambridge: MIT Press 351-365.

Dretske, F. 1993. Can Intelligence be Artificial? *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 71(2): 201-216, August 1993.

Dowe, D. & Oppy, G. 2016. *The Turing Test. Stanford Encyclopaedia of Philosophy.* [Online] Available: http://plato.stanford.edu/entries/turing-test/ [8 October 2016].

Edmonds, B. 2000. The Constructability of Artificial Intelligence (As Defined by the Turing Test). *Journal of Logic, Language and Information*, 9(4): 419-424, October 2000.

Frowe, H. 2009. The Justified Infliction of Unjust Harm. *Proceedings of the Aristotelian Society*, 109: 345-351.

Future of Life Institute: Open Letter on Autonomous Weapons. [Online]. Available: http://futureoflife.org/open-letter-autonomous-weapons/ [28 April 2016].

Gayle, D. 2016. *Police with Body Cameras Receive 93% Fewer Complaints – study*. [Online] Available: https://www.theguardian.com/uk-news/2016/sep/29/police-with-body-cameras-receive-93-fewer-complaints-study [27 February 2017].

Globalsecurity.org: Counter Rocket, Artillery and Mortar (C-RAM). [Online] Available: http://www.globalsecurity.org/military/systems/ground/cram.htm [26 February 2017].

Gormley, D.M. 2009. *Winning on Ballistics but Losing on Cruise: The Missile Proliferation Battle.* [Online] Available: https://www.armscontrol.org/act/2009_12/Gormley#Table1 [26 September 2016].

Harris, S. 2012.*Out of the Loop: The Human Free Future of Unmanned Aerial Vehicles.* [Online]. Available: http://www.hoover.org/sites/default/files/research/docs/emergingthreats_harris.pdf [02 August 2016].

Henley, T.B. 1990. Natural Problems and Artificial Intelligence. *Behaviour and Philosophy*,18(2): 43-56, Fall/Winter 1990.

Hodges, A. 2013. *Alan Turing. Stanford Encyclopaedia of Philosophy.* [Online] Available: http://plato.stanford.edu/entries/turing/#TurMacCom [18 May 2016].

Human Rights Watch: Shaking the Foundation: Human Rights Implications of Killer Robots. [Online] Available: https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-implications-killer-robots [17 July 2016].

International Campaign to Ban Landmines: The Treaty: Treaty Status. [Online] Available: http://icbl.org/en-gb/the-treaty/treaty-status.aspx [16 May 2016].

International Committee of the Red Cross, 2009. *Interpretive Guidance on the Notion of Direct Participation in Hostilities Under International Humanitarian Law*. [Online] Available: https://www.icrc.org/eng/assets/files/other/icrc-002-0990.pdf [Accessed 24 August 2016].

International Committee of the Red Cross, 2016. *Convention on Certain Conventional Weapons: Meeting of Experts on Lethal Autonomous Weapons Systems*. [Online] Available: https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system [24 August 2016].

Jacob, P. 2014. *Intentionality. Stanford Encyclopaedia of Philosophy*. [Online] Available: http://plato.stanford.edu/entries/intentionality/ [2 October 2016].

Johnson, D.G. 2014. *Technology with No Human Responsibility? Journal of Business Ethics.* [Online] 127, (4), pp. 707-715. Available:

https://www.law.upenn.edu/live/files/3774-johnson-d-technology-with-no-responsibility [16/03/2016].

Lazar, S. 2016. *War.* Stanford Encyclopaedia of Philosophy. [Online] Available: http://plato.stanford.edu/entries/war/#JusAdBell [19 July 2016].

Levin, J. 2016. *Functionalism.* Stanford Encyclopaedia of Philosophy. [Online] Available: http://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=functionalism [09 October 2016].

McCoy, T. 2014. *A Computer Just Passed the Turing Test in a Landmark Trial*. The Washington Post. [Online] Available: https://www.washingtonpost.com/news/morning-mix/wp/2014/06/09/a-computer-just-passed-the-turing-test-in-landmark-trial/ [5 October 2016].

O'Connell, M.E. 2014. Banning Autonomous Killing, in: Evangelista, M. and Shue, H. (eds.). *The American Way of Bombing: Changing Ethical and Legal Norms, From Flying Fortresses to Drones*. Cornell: Cornell University Press. 224-234.

Ohlin, J.D. 2016. *The Combatant's Stance: Autonomous Weapons on the Battlefield*. [Online] Available: https://www.law.upenn.edu/live/files/3916-ohlin-jens-machine-liability-and-the-combatants [16 July 2016].

Searle, J.R. 1980. Minds, Brains and Programs. *The Behavioural and Brain Sciences*, 3(3): 417-424.

Sharkey, N. 2012. The Human Control of Weapons: A Humanitarian Perspective. Draft to appear in: Beck et al, S. Autonomous Weapons Systems: Law, Ethics, Policy (in press).

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

Turing, A.M. 1950. Computing Machinery and Intelligence. *Mind*, 59(236): 433-460, October 1950.

United States Department of Defence, 2012. *Directive Number 3000.09.* Arlington: Department of Defence. P. [Online] Available: http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf [25 May 2016].

Weizenbaum, J. 1966. ELIZA – A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the Association of Computing Machinery*. 9(1): 36-45.