Autonomous Weapon Systems and Meaningful Human Control

University of Oulu
Department of Information Processing Science
Bachelor's Thesis
Pauli Klemettilä
14.6.2022

Abstract

Autonomous weapon systems are an emerging technology that affects the lives of soldiers and civilians who operate or live in conflict zones. The ethical and humanitarian problem caused by the advent of artificial intelligence -based weapon systems capable of independently making decisions about target selection and engagement has led to substantial academic debate, with critics calling for a total ban on such weapons based on the grounds that their use will invariably lead to negative humanitarian consequences, while others have suggested that autonomous systems may, in some situations, be able to behave in a more humane fashion on the battlefield than human soldiers.

This thesis analysed the current state of research on the topic of autonomous weapon systems through a literature review. The concepts of autonomous weapon system and meaningful human control were defined and explained. Arguments about the humanitarian consequences of developing these weapon systems, both positive and negative, were critically examined. A special focus was placed on examining the proposed methods and principles for designing autonomous weapon systems in a way that mitigates their potential negative humanitarian effects.

The results of the literature review indicated that the concept of meaningful human control is the most prominent and promising principle to guide the design and development of autonomous weapon systems that adhere to the international humanitarian law. Value sensitive design was found to be a potential methodology for injecting human values into the design process of autonomous weapon systems.

Keywords
Autonomous weapon systems, meaningful human control, artificial intelligence, robotics, warfare

Supervisor

University teacher Jouni Lappalainen

Contents

Abstract 2
Contents 3
1. Introduction 4
2. Research Method 6
3. Central Concepts and Previous Research 7
   3.1 Definition and importance of autonomous weapon systems 7
   3.2 International humanitarian law 9
   3.3 Criticism of autonomous weapon systems 10
   3.4 Meaningful human control 11
   3.5 Proposed humanitarian benefits of autonomous weapon systems 12
   3.6 Potential methods and guidelines for designing ethically compliant autonomous weapon systems 13
4. Analysis 15
5. Discussion 17
6. Conclusions 19
References 21

1. Introduction

The concept of an artificial intelligence breaking free from human control and wreaking havoc upon or even destroying humanity has been a frequent theme in science fiction and futuristic speculation for decades. Movies such as *Terminator* and *2001: Space Odyssey* are engraved in the public consciousness, contributing to fears about an artificial intelligence revolution (LeCun & Zador, 2019). Public figures such as Elon Musk and Stephen Hawking have warned against uncontrolled AI development as a potential existential threat (Cellan-Jones, 2014). However, it has been argued that such fears are unfounded, as artificial intelligence ultimately serves whatever goals humans have set for it (LeCun & Zador, 2019). Whether or not the apocalyptic vision of a robot revolution is a real possibility, there are far more contemporary and realistic dangers and ethical issues surrounding the development and deployment of autonomous systems that make decisions affecting the wellbeing of humans. One of the applications of AI where these problems are markedly present is the field of autonomous weapon systems.

Hynek and Solovyeva (2021) define autonomous weapon systems by a unique set of attributes: they are fully autonomous and able to engage in decision-making, they can be used as offensive weapons, and their existence has been made possible by advances in AI technology. Machine autonomy with respect to life-and-death decisions is the most intensely debated aspect of these technologies, with critics calling for a global ban on their development (Hynek & Solovyeva, 2021). Wallach (2017) proposes that machines should never be the ones making such decisions, as they are not moral agents. Additionally, he argues that it is difficult to test these systems adequately in order to ensure that they behave predictably in complex scenarios. However, Arkin (2013) argues that suitably deployed robotics technology could lead to reduction in civilian and combatant deaths, as many atrocities of war can often be attributed to flaws in human behaviour that do not necessarily apply to machines, such as the tendencies to dehumanize or seek revenge on the enemy.

This thesis is a literature review in which the following research questions are explored:

What are the consequences of giving AI-based weapon systems the capability to make life and death decisions?

Could these systems be designed, developed and tested thoroughly enough so that their use in place of human soldiers would either mitigate the atrocities of war or at least cause no increase in them?

Methods presented in previous research that propose to avoid or mitigate the negative consequences of autonomous weapon systems are critically examined, with a particular focus on the concept of "meaningful human control over autonomous systems" (Santoni de Sio, 2018). For the purposes of this thesis, the goals of international humanitarian laws relating to armed conflict as presented in *the Practical Guide to Humanitarian Law* (Bouchet-Saulnier, 2013) published by Doctors without Borders serve as a basis for the

ethical framework against which the attributes of autonomous weapon systems are compared.

At the time of writing this thesis the debate on autonomous weapon systems is more relevant than ever, as a United Nations panel reported in 2021 that an autonomous weapon system was deployed in combat for the first time ever in a March 2020 skirmish in Libya (Russell et al., 2021). This incident has been described as a landmark moment, demonstrating that the "red line" of autonomous targeting of humans has been crossed (Russell et al., 2021). More recently, there have been reports that Russia used a loitering munition capable of AI-based real-time target recognition and classification of detected objects against Ukraine in March of 2022 (Kallenborn, 2022).

The goals of this thesis are to compare the arguments for and against the development of autonomous weapon systems and to examine possible ways to find common ground between them by investigating if these systems could be designed and developed to behave on the battlefield in ways that are equally or more compliant with humanitarian laws when compared to the behaviour of human soldiers. Finding some sort of compromise between an outright ban and complete freedom in the use and development of these weapons is highly important, for it is unlikely that all nations of the world would agree to such a ban, as is the case with nuclear weapons.

As this thesis is a literature review, all discussion is based on what has been previously published on the topic of autonomous weapon systems. As such technologies are highly classified by states and companies that develop them, the analysis is primarily based on expert opinions on what the capabilities of current autonomous weapon systems are and how these capabilities may develop in the future. Additionally, ethical questions about weapons research in general are not discussed in this thesis.

This thesis is structured into six chapters including this introduction. The following chapter goes into more detail about the research method used in this thesis (literature review). Chapter three presents the previous research and central concepts that form the basis for this literature review. In chapter four, the reviewed research and its relevance is analysed. Chapter five discusses the findings of this literature review and their implications. Chapter six presents the conclusion and suggests further areas of research pertaining to the subject of autonomous weapon systems.

2. Research method

This thesis is a literature review examining and comparing the main arguments of scientific papers debating the issue of autonomous weapon systems in order to answer the previously presented research questions. Literature review is a form of study that summarizes and evaluates the state of knowledge or practice on a specific subject by examining a collection of relevant writings (Knopf, 2006). Traditionally literature reviews focus on academically published books and articles, but actors that are not university-based, such as non-governmental organizations and think tanks, might also issue reports that are relevant (Knopf, 2006). By identifying and comparing the theories and evidence presented in reviewed studies it is possible to identify the underlying disagreements that are the root cause for debates in the literature (Knopf, 2006). This makes literature review an appropriate choice of study method for evaluating the current state of the academic debate about autonomous weapon systems.

Most of the papers used for this review come from the field of computer science and more specifically robotics and artificial intelligence. However, the discussion surrounding autonomous weapons is multidisciplinary in nature, and therefore papers from the fields of law, philosophy and military studies are included as well. Furthermore, reports on autonomous weapons authored by United Nations and other humanitarian organizations are likewise taken into account.

The literature used for this review was primarily found by searching Google Scholar and the databases ACM Digital Library and IEEE Xplore. Search terms used included “autonomous weapon systems”, “meaningful human control”, “killer robots” and “international humanitarian law.” It is notable that most of the scientific literature discussing the subject of autonomous weapon systems appears to be critical of the concept, so arguments to the contrary were specifically searched for by using terms and phrases such as “humanitarian benefits”, “benefits”, “in defence of” and “positive effects” in conjunction with the previously mentioned terms, using combinations of AND/OR operators when applicable.

As autonomous weapon systems are an emerging technology, most of the literature used in this thesis was published during the last five years. The oldest citation is from 2008, while the newest scientific source used was published in December of 2021. Recentness was one of the criteria on which the papers used were selected.

As most of the included papers do not include empirical research, instead often being argumentative essays presenting different viewpoints on the topics of autonomous weapon systems and meaningful human control, the selection of the studies was primarily based on an evaluation of whether their arguments are relevant to this study and their influence in terms of academic citations. The arguments of the cited works are analysed and compared against each other in an effort to find answers to the research questions.

The overall frame of reference from which these studies are looked at is humanitarian harm reduction in the context of designing and developing autonomous weapon systems. It is optimistically assumed that the actors involved in developing and deploying these weapon systems wish to design them in such a way as to make it possible to keep civilian casualties and other unnecessary harm they cause to a minimum. However, it is acknowledged that these actors might not necessarily be willing to sacrifice strategic or tactical benefits of these weapons for the purposes of harm reduction.

3. Central Concepts and Previous Research

This chapter introduces the concepts and previous research that are essential to this thesis, starting with finding a definition for autonomous weapon systems and establishing that they are, in fact, an already existing and relevant phenomenon. The primary arguments that are against the development of autonomous systems are explained, and the concept of meaningful human control is introduced. Arguments from previous research claiming that autonomous weapon systems could in fact have positive humanitarian effects on warfare are presented, along with proposed methods and design principles to guide the development of these systems in order to achieve these positive results, or at least mitigate negative ones.

3.1 Definition and importance of autonomous weapon systems

By its most common definition, an autonomous weapon system (AWS) is a machine than can identify enemy targets and perform lethal actions in combat situations without direct human input (Charters, 2020). In public discourse, autonomous weapon systems are commonly referred to with terms such as “slaughterbots” (Russell et al., 2021) or “killer robots” (Hynek & Solovyeva, 2021). While such terms may conjure up images of humanoid robots from science fiction, the most typical real-world example of an autonomous weapon system is an AI-based military drone (Piper, 2019).

Autonomous weapon systems are rapidly becoming a reality due to improvements in machine learning, robotics and automation in general (Righetti et al., 2018). Weapon systems incorporating artificial intelligence may even be capable of independently learning how to improve how they conduct targeting (Petman, 2017).

The primary reasons for developing autonomous weapon systems are the military advantages they provide. Autonomous weapon systems reduce the number of soldiers required for a given mission and enable expanding the battlefield to areas that are inaccessible to human soldiers (Ezioni, 2017, as cited in Felt, 2020). Unlike humans, autonomous weapons are not prone to errors that are caused by getting tired, frustrated or over-confident (Kallenborn, 2021). Additionally, maintaining a single human combatant can be significantly more expensive than a small rover equipped with weapons (Etzioni, 2017, as cited in Felt, 2020). Autonomous weapon systems, such as drone swarms, can also easily be deployed on “suicide missions” with the assumption that they do not return.

There are varying degrees of autonomy, and some weapon systems can switch between an autonomous and a non-autonomous state (Charters, 2020). In theory, even simple weapons such as landmines can be argued to have some level of autonomy, as their destructive powers are not directly controlled by humans once they have been laid (Petman, 2017). A higher level of autonomy is possessed by weapons such as sentry guns that automatically fire at targets within range and systems that automatically detect and fire at incoming missiles (Petman, 2017). However, none of these are considered truly autonomous systems, as they do not incorporate strong artificial intelligence with decision-making capabilities (Petman, 2017).

While robotics often makes no distinction between autonomous and automated systems, the separation is important in the context of differentiating between types of weapons

(Righetti et al., 2018). Even though there are already existing weapon systems that can independently identify and fire at enemy combatants or projectiles based on pre-programmed algorithms, a weapon system is truly autonomous only if it is capable of independently performing a deliberative process of judgement that determines when and against whom to employ lethal force (Caron, 2020).

In order to clearly differentiate autonomous weapon systems from non-autonomous weapons that can cause damage while not necessarily being supervised by humans, a specific definition is required. Wyatt (2020) argues that finding a concrete definition for what constitutes an autonomous weapon system is one of the major stumbling blocks to developing an international response to the emergence of autonomous military technology. As mentioned in the introduction, according to Hynek and Solovyeva (2021) autonomous weapon systems can be differentiated from other weapon categories by possessing the following set of attributes:

1. They are fully autonomous and able to engage in lethal decision-making, targeting and use of force.

2. They can be used as offensive weapons in dynamic environments, possibly without direct human control or supervision.

3. Advances in AI are what characterizes these systems and what has made their development possible.

This list of defining attributes given by Hynek and Solovyeva is not entirely without issues, as the question of whether a weapon system should be considered autonomous is not necessarily a binary one between full autonomy and full human control. According to a model published by Human Rights Watch (2012, as cited by Alwardt & Kruger, 2016), robots which can select targets and deliver force under the supervision of a human operator who can override their actions are not fully autonomous but are still categorized as autonomous weapons. For the purposes of this thesis, any AI-based weapon system with the capability to identify and engage targets on its own without direct human input is considered to be in the category of autonomous weapon systems, even if human operators still possess the ability to veto its actions.

Due to a multitude of reasons, such as maintaining technological advantage over other nations or economic competitors, weapons technology is one of the most secretive fields of technological research. It is not publicly known what types of autonomous weapons are currently either in development or ready for deployment. However, weapons which have both remote-control and autonomous capabilities have already been used in combat, even if it is unclear whether humans have made the final decisions about attacking individual targets (De Vynck, 2021; Hernandez, 2021).

At the time of writing this thesis, the discussion surrounding autonomous weapon systems is, for the most part, forward-thinking, as fully autonomous systems are not yet widely used by the world's militaries (Charters, 2020). However, as remarked in the introduction, according to a 2021 U.N. report, a drone airstrike that was conducted in Libya in the spring of 2020 was carried out by Turkish Kargu-2 drones which are capable of fully autonomous targeting, can remain operational without GPS or radio links and are

equipped with facial recognition software for targeting humans (Russell et al., 2021). According to Russell et al., this demonstrates that the age of autonomous weapon systems is already here. It must, however, be noted that it is not entirely clear whether this drone was operating autonomously at the time of the attack (Hernandez, 2021; Kallenborn, 2021). Nevertheless, it is notable that weapons with capability to act fully autonomously have demonstrably been used in combat situations. The question of what the implications of autonomous weapon systems are in the context of international humanitarian law has therefore become extremely relevant.

3.2 International Humanitarian Law

The term international humanitarian law (IHL) refers to the branch of public international law that concerns the laws of armed conflict. Its purpose is to reduce suffering in warfare by governing how hostilities are conducted (Bouchet-Saulnier, 2013). The international humanitarian law is one of the oldest laws in existence, having been revised multiple times throughout history by a series of treaties and agreements between states (Bouchet-Saulnier, 2013).

According to the goals of international humanitarian law, means and methods of warfare should be restricted in order to avoid unnecessary suffering and destruction (Bouchet-Saulnier, 2013). Additionally, it posits that non-combatants should be protected from the destruction caused by war (Bouchet-Sailnier, 2013). The United Nations has affirmed that international humanitarian law applies to autonomous weapon systems, but meetings between states discussing a pre-emptive ban of the development of autonomous weapon systems have shown that there is a substantial opposition among various states to such a ban (Evans, as cited by Felt, 2020).

Autonomous weapon systems present a regulatory challenge for international humanitarian law, as they replace the human role as the killer in war (Petman, 2017). Additionally, these weapons must satisfy multiple requirements in order to be compatible with international humanitarian law. For example, they must be able to distinguish between civilian and military targets, as one of the core tenets of the IHL is the distinction between non-combatants and combatants (Bouchet-Sailnier, 2013).

The three core principles of international humanitarian law that are the most relevant to autonomous weapon systems are distinction, proportionality, and precaution (Righetti et al., 2018). The principle of distinction is the previously mentioned requirement to always distinguish between combatants and civilians. The principle of proportionality exists to set boundaries for the amount and type of force that can be used in a conflict in order to prevent excessive damage to civilians and civilian objects. The principle of precaution states that the participants of a conflict must take all precautions to protect civilians who are under their control from the effects of attacks (Righetti et al., 2018).

The question whether autonomous weapon systems can be designed to function in such a manner as to comply with the principles of international humanitarian law lies at the heart of the debate presented in this thesis, as some critics (Wallach, 2017) have argued that autonomous weapon systems are fundamentally violating the principles of international humanitarian law, while others have claimed that they can in fact be superior to human-operated systems in following these principles (Charters, 2020).

3.3 Criticism of autonomous weapon systems by academics and organizations

In 2017, the Future of Life Institute released a video dramatizing the potential consequences of widespread use of autonomous weapon systems titled “Slaughterbots” (Russell et al., 2021). An open letter from over 3,100 researchers from the fields of robotics and artificial intelligence have signed an open letter advocating a pre-emptive ban on the development of autonomous weapon systems (Wallach, 2017). A survey of attitudes towards autonomous weapon systems targeted to machine learning and artificial intelligence scholars found that 52% strongly opposed other researchers working on autonomous weapon systems and 42% would either resign or threaten to resign if their university supported such research (Zhang et al., as cited by Wyatt & Galliott, 2021). While a complete ban on the development and deployment of autonomous weapon systems may or may not be possible due to political factors, the major criticisms against them could at the very least serve to create a set of boundaries that would guide their development.

Practical problems involving autonomous weapon systems include the fact that it can be difficult to stop them from ending up in the hands of non-authorized actors, as it may be possible to assemble them using technologies that have been developed for civilian applications (Wallach, 2017). While leading military powers have claimed that they will maintain effective human control over their deployed autonomous weapon systems, they cannot guarantee that other state or non-state actors will do the same, even if their sincerity is accepted at face value (Wallach, 2017). Additionally, it has been suggested that autonomous weapon systems will inevitably evolve into weapons of mass destruction, as they can be deployed in great numbers (i.e., drone swarms) without human supervision (Russell et al., 2021). Autonomous weapons are also potentially vulnerable to cybersecurity threats (Righetti et al., 2018), with the worst-case scenario being takeover by terrorist groups.

It has been argued that autonomous weapon systems may never be able to distinguish civilian and military targets on a sufficient level (Human Rights Watch, 2015), which would be against the IHL principle of distinction (Righetti et al., 2018). It has been called into question whether it is possible for an algorithm to find a reasonable balance between expected civilian casualties and military advantage (Righetti et al., 2018).

Concerns have been raised about the predictability and testability of autonomous weapon systems (Righetti et al., 2018; Wallach, 2017). In the case of weaponry, predictability means that “within the task limits for which the system is designed, the anticipated behavior will be realized, yielding the intended result” (Wallach, 2017). Doubts have been raised about whether an autonomous weapon system can be expected to assess every situation correctly and predictably, especially when conditions become unexpected and extremely adversarial (Righetti et al., 2018). According to Wallach (2017), unpredictable behaviour from autonomous weapons will occasionally result in non-combatant deaths, ignite new conflicts or escalate hostilities.

Wallach (2017) argues that reasonable testing procedures for autonomous weapon systems are not exhaustive and therefore will not ensure that these complex and adaptive systems will behave predictably on the field of battle. Furthermore, each software update can have such effects on the behaviour of autonomous weapon systems that additional rounds of extensive testing would be required, which may be ignored by states under pressure to cut military expenditures (Wallach, 2017). Weapon systems capable of

machine learning on the battlefield are even more problematic from this perspective (Wallach, 2017).

Autonomous weapon systems may also be more vulnerable to outside interference than human soldiers. Klincewicz (2015) argues that because AWS software is likely to have unobserved bugs, these systems are susceptible to hacking. In the worst-case scenario, autonomous weapon systems could even be taken over by malevolent entities such as terrorist or criminal organizations (Klincewicz, 2015). In conflict situations, belligerents may try to damage the sensors and data receivers of autonomous weapon systems or deceive them by feeding them false data, causing them to fail in unpredictable ways (Holland Michel, 2021).

Critics of autonomous weapon systems often highlight the fundamental ethical issues that arise from the removal of direct human involvement from the decision to use lethal force (Wyatt & Galliott, 2021). It has been argued that only moral agents – beings able to reason about what is right and what is wrong – should be able to make life and death decisions (Wallach, 2017; Amoroso & Tamburrini, 2020). While such arguments do have merit from a moral perspective, one of the more empirical issues of giving machines this power is the accountability gap that is potentially created.

In the case that an autonomous weapon system makes targeting decisions that are against the IHL (i.e., decides to shoot and kill non-threatening civilians), it can be difficult to determine personal accountability for these crimes (Amoroso & Tamburrini, 2020). While the issues of accountability in unclear combat situations are not exclusive to autonomous weapons, the issue becomes complex when there is no direct causal link between the use of force and a human who can be held accountable for the decision (Wyatt & Galliott, 2021).

3.4 Meaningful Human Control

Meaningful human control (MHC) is a concept that broadly refers to human participation in and contribution to the operation of artificial intelligence systems, especially when these systems are deployed to operate autonomously in high-stakes situations (McCoy et al., 2019). Roff et al. (2016) argue that human control over technology, including but not limited to autonomous weapon systems, is enhanced if the technology is predictable, reliable, transparent and provides the user with accurate information.

At a basic level, the need for meaningful human control arises from the premise that a machine operating without human control of any kind is widely considered unacceptable (Roff et al., 2016). Humans are adaptable and able to respond to unusual situations, while machine-learning based systems that have been trained on predetermined sets of inputs may perform poorly when confronted with novel or atypical inputs (McCoy et al., 2019). As mentioned previously, an autonomous system performing undesirable actions without direct human control may result in an accountability gap where it is difficult to determine who is at fault for the events in question (Amoroso & Tamburrini, 2020).

The concept of meaningful human control is most often brought up in the literature on autonomous weapon systems (McCoy et al., 2019). According to this principle, humans should ultimately remain in control of, and morally responsible for, decisions about lethal military operations as opposed to giving this power to autonomous machines (Santoni de Sio & van den Hoven, 2018).

While there is no universally agreed upon definition of meaningful human control (Santoni de Sio, 2018), various thresholds that would have to be met in order for human control to be meaningful have been proposed. According to Amoroso and Tamburrini (2020), for human control over an autonomous weapon system to be meaningful, the system must satisfy the following three requirements:

1. Human control must act as a fail-safe that prevents a malfunctioning weapon from attacking civilians or causing excessive collateral damage.

2. Human control must ensure that responsibility can be ascribed to the right individual in the case that an autonomous weapon commits unlawful acts.

3. Human control should ensure that moral decisions affecting the life, physical integrity and property of people involved in armed conflicts are not made by artificial agents.

Meaningful human control is inherently a concept that is not without issues, because while definitions such as the one given above exist, other definitions for what makes human control meaningful in the context of autonomous weapon systems "range from a military leader specifying a kill order in advance of deploying a weapon system to having the real-time engagement of a human in the loop of selecting and killing a human target" (Wallach, 2017). These problems in concretely defining such concepts as meaningful human control have repeatedly surfaced in U.N. discussions about autonomous weapon systems as well (Wallach, 2017). In this thesis the above attributes given by Amoroso and Tamburrini are used in order to define the minimum requirements that must be satisfied in order for human control over an autonomous weapon system to qualify as meaningful.

3.5 Proposed humanitarian benefits of autonomous weapon systems

While it can be counterintuitive to consider that development of new weapons technology might lead to positive humanitarian consequences, arguments have been made that the development of autonomous weapon systems could potentially lead to such results. Arkin (2013) argues for the possible benefits of autonomous weapon systems on the basis that the historical and current track records of ethical behaviour on the battlefield are not favourable to humans, asserting that suitably deployed robotics technology could reduce non-combatant deaths and casualties. According to Arkin (2013), there are several reasons why autonomous weapon systems could potentially be better at adhering to international humanitarian law on the battlefield than human soldiers.

Arkin (2013) proposes that because autonomous weapon systems do not need to have self-preservation as a leading priority in situations where target identification is uncertain, they do not have to operate on the principle of 'shoot first and ask questions later' and can therefore utilize an approach where they do not initiate hostilities, only respond to them. Assuming risks in order to prevent civilian casualties can therefore be easier for machines than human soldiers (Arkin, 2013).

According to Arkin (2013), intelligent autonomous systems can quickly gather information from sources such as remote sensors and human intelligence before deciding if the use of lethal force is required, while increasingly network-centric modern warfare will progressively become too difficult for humans to direct. Therefore, an autonomous weapon system could possibly make the decision to hold fire in a situation where human soldiers would unnecessarily decide to perform hostile actions due to having incomplete information. Robotic systems could also potentially be able to pierce the fog of war more effectively than humans (Arkin, 2013), which could, for example, reveal that there are civilians present in the combat zone. A combined team of human soldiers and autonomous systems could also be able to effectively monitor the ethical behaviour of all participants on the battlefield, gathering evidence of infractions and reporting them (Arkin, 2013).

Arkin (2013) points out that autonomous weapon systems are not subject to human emotions or cognitive biases which can contribute to poor decision making on the battlefield. Anger, fear, frustration and refusal to accept incoming contradictory information in stressful situations all contribute to undesirable human behaviour and poor decision-making on the battlefield (Arkin, 2013). Charters (2020) argues that autonomous weapons "do not kill at random or for pleasure", instead being able to base their decisions on international laws that are hard coded into their software.

Dunlap (2016) claims that the previously mentioned arguments against autonomous weapon systems based on the issue of accountability are without merit from a legal perspective if these weapons have been tested and designed so that their expected actions can be reasonably predicted, as any weapons of war can sometimes function in unintended ways despite rigorous testing, and laws about involuntary manslaughter can be applied to situations where autonomous weapons are used negligently.

3.6 Potential methods and guidelines for designing ethically compliant autonomous weapon systems

If there is to be any hope of developing autonomous weapons that could be used on the battlefield with confidence that they will act in accordance with international humanitarian laws, they must be thoroughly tested under realistic conditions before deployment (Petman, 2017). Because safety and security critical autonomous systems are expected to be able to perform in unpredictable situations and contexts, it is fundamentally impossible to test them for a specified subset of such situations (Song et al., 2021). According to Song et al., traditional brute force testing cannot be scaled for autonomous systems, and therefore new guidelines and approaches for testing are required. According to Song et al. (2021), having access to realistic datasets is of paramount importance for both training and testing of autonomous systems. Additionally, testing these systems requires an iterative and continuous approach to engineering (Song et al., 2021).

Song et al. (2021) examine various testing techniques and approaches that could be applied to autonomous systems. Model-based testing approaches could be utilized to model the properties, constraints and behaviour of autonomous systems and could be further used for automatically generating and executing test cases (Song et al., 2021). Additionally, formal methods can be used to "analyse and represent the system design, inputs and outputs using domain terminologies, and validate the system against a formal specification" (Song et al., 2021). However, most current academic contributions to the testing of autonomous systems are based on adaptations of approaches used for testing conventional systems, and autonomous systems do require novel approaches to testing

(Song et al., 2021). Therefore, while these approaches may well be adaptations and combinations of techniques used for conventional systems, research should be conducted in order to develop techniques that could be used for autonomous systems in general (Song et al., 2021). If such techniques were developed, they could potentially also be used for developing autonomous weapon systems that comply with the international humanitarian law.

Arkin (2008) suggested that autonomous weapon systems should be programmed to perform a series of gated decisions based on the IHL principles of distinction and proportionality before being able to use lethal force. Wyatt and Galliott (2021) have proposed a framework for guiding the development of autonomous weapon systems based on value sensitive design, a methodology that seeks to inject human ethics into an iterative design process. It is based on the premise that misuses of technology can be minimised by integrating positive ethics into the development process at the design stage. De Sio and van den Hoven (2018) also promote value sensitive design as an ideal to guide designing autonomous weapon systems that are under meaningful human control.

The value sensitive design process consists of three stages: a conceptual investigation where direct and indirect stakeholders and potential harms or benefits to them are identified, an empirical evaluation of possible stakeholder experiences and social impacts of interacting with the technology, and technological investigation where it is determined how the technology supports or constraints human values, and how the values identified in previous stages could be included to the design process (Wyatt & Galliott, 2021). In the context of autonomous weapon systems, it is required to consider the perspectives of both civilian and military stakeholders (Wyatt & Galliott, 2021). Wyatt and Galliott note that even military stakeholders have ranked reduction in the risk of harm to civilians as a matter that is almost equally important as the risk to service personnel when considering the matter of autonomous weapon systems.

Arkin et al. (2019) propose a middle road between an outright ban and uncontrolled development of autonomous weapon systems, stressing the importance of developing technological measures in order to mitigate the possibility of autonomous weapon systems causing unintentional escalation in conflict situations. Examples of this include making sure that autonomous weapons operate under a no-first-fire policy and therefore do not initiate hostilities, and that their missions are always directly provided to them by humans (Arkin et al., 2019). Additionally, communication links used by autonomous weapon systems must be resilient in order to ensure that they can be recalled by humans in the event of unauthorized behaviour (Arkin et al., 2019).

As autonomous weapon systems are a futuristic technology, further research into improving relevant technologies and human-machine systems should be conducted in order to ensure IHL compliance of these systems and to reduce the non-combatant harm they cause (Arkin et al., 2019). New methodologies and techniques for ensuring the reliability and security of these systems should be developed, and strategies to promote human participation in decision-making about the use of force should be improved (Arkin et al., 2019).

4. Analysis

The analysis of previous research indicates that autonomous weapon systems are indeed currently in development and are to some extent an already existing technology. While defining what constitutes an autonomous weapon system is not a straightforward task, it is possible to find a working definition by basing it on their capabilities for independent deliberation and decision making when it comes to targeting and the use of lethal force. However, the degree of autonomy required for a weapon system to qualify as autonomous remains a matter of debate in the literature.

International humanitarian law provides constraints for the development of autonomous weapon systems, as these machines must comply with its principles of distinction, proportionality and precaution in warfare. While many organizations and a substantial amount of academics from the fields of AI and robotics research heavily speak against the development of these weapons, it appears that a global U.N.-mandated ban of their development is not in the horizon due to opposition against such a ban by leading military powers. Therefore, it is important to find ways to ensure that the development and deployment of autonomous weapons does not result in an increase in civilian casualties and other wartime atrocities.

Surveys conducted in previous research indicate that opposition against the idea of autonomous weapon systems is the prevalent position of academics who work in the field of robotics and artificial intelligence. The main criticisms against these weapon systems either highlight the inherent ethical problem of giving machines the power to make life and death decisions or the more practical issues of predictability, testability, reliability, vulnerability and accountability. Some of the criticisms directed at autonomous weapon systems, such as the risk of them falling into the wrong hands, could be applied to any weapons technology, autonomous or otherwise. However, the autonomous nature of these weapons clearly does lead to specific issues being highlighted.

While the issues of predictability, reliability, vulnerability and testability are not exclusive to autonomous weapon systems, the absence of a human operator leads to them becoming especially important, as there is potentially no one performing corrective actions in the case of malfunctions. Furthermore, the argument about the dangers of autonomous weapon systems being hacked by outside actors is particularly compelling, because an autonomous weapon gone rogue might be able to cause a substantial amount of destruction before being destroyed or deactivated if there is no human operator authorizing its individual decisions.

A similar issue highlighted in previous research is the matter of accountability gap. Unclear situations regarding accountability when it comes to war crimes are not uncommon in conventional warfare but adding in the factor of weapons that autonomously make decisions about using lethal force undeniably can lead to the addition of another layer of difficulty in bringing those responsible to justice, even though the assertion that these weapon systems should be banned on these grounds has been challenged from a legal perspective. Overall, previous research criticizing autonomous weapon systems does establish that the ethical position which considers giving autonomous weapons the capability to independently make life and death decisions morally unacceptable is supported by practical concerns raised by these systems.

Based on the conducted searches and the reviewed literature, it can be stated that academic views emphasizing the positive humanitarian benefits of autonomous weapon

systems are in the minority. The primary arguments that support the possibility of these benefits are based on the notion that machines are not subject to human emotions, which are often contributors to unethical human behaviour on the battlefield, and on the fact that unlike humans, autonomous weapon systems do not have to have self-preservation as an utmost priority and therefore they can be programmed to follow a policy of never firing the first shot. There is also the debatable argument that machines can process incoming information more quickly and objectively than humans, therefore being able to make more informed decisions in the heat of the moment on the battlefield.

It should be noted that achieving many of the potential humanitarian benefits of autonomous weapon systems require limiting their capabilities in ways that might reduce their strategic effectiveness. For example, an autonomous weapon that only responds to offensive actions and does not initiate them is clearly at a higher risk of being destroyed before being able to complete its mission. Whether or not implementing such attributes to autonomous weapon systems is realistic depends on what kind of global regulations will be imposed upon them and on the interests of stakeholders. This is where the views of the cautious proponents and the critics of autonomous weapon systems start to have some overlap, as without suitable guidelines to guide their development from early on, the development of autonomous weapon systems will have negative consequences from a humanitarian perspective. An example of this overlap is the previously cited workshop paper from Arkin et al. (2019), where academics with different and in some cases divergent views on autonomous weapon systems come together to explore ways of dealing with the problems that arise from the development of these weapons.

While many of the solutions to the issue of autonomous weapon systems proposed in the analysed literature are regulatory in nature and therefore to some extent dependent on legislative bodies, some of the problems could potentially be solved with techniques that are used for the design and development of autonomous systems in general. The vital question of whether it is possible to develop sufficient testing procedures for ensuring the reliability and security of autonomous weapon systems was highlighted in several of the reviewed studies, including Petman (2017) and Wallach (2019). Testing of autonomous systems provides challenges that are not present in testing conventional systems, especially when the autonomous systems are both safety and security critical, which is a description that certainly applies to autonomous weapon systems. As mentioned before, Song et al. (2021) suggest that the current state of research into testing autonomous systems in academia and industry is in need of improvement. However, various testing approaches and techniques, such as model-based testing and formal methods, do show promise for testing autonomous systems.

The most prominent proposed solution to the issue of autonomous weapon systems presented in previous research is the concept of meaningful human control. While the analysed literature makes it clear that there is no universally accepted definition of meaningful human control, it can be broadly defined as a set of constraints to prevent these weapons from acting in a completely autonomous capability and to ensure that human operators remain ultimately responsible for the actions conducted. An adjacent issue to finding a concrete definition for meaningful human control is the matter of finding a suitable quantitative threshold for what makes human control meaningful. According to previous research, value sensitive design could be utilized for injecting meaningful human control into the design process of autonomous weapon systems from the beginning.

5. Discussion

In order to answer the research questions of this thesis, several issues highlighted by the analysed literature need to be discussed. The question of finding a concrete definition for what level of autonomy qualifies a weapon as an autonomous system is vital for any sort of analysis of the consequences of the development and deployment of these weapons. By focusing on the quality of having the capability for an independent, deliberative processes of target selection and lethal decision making, it is possible to arrive at a definition that both facilitates academic discussion and provides to the public discourse a non-sensationalised view of what autonomous weapons realistically are. Even though a concrete definition is important, in some cases it may be difficult to determine where the wavering line between autonomous and automated weapons is drawn, especially if the military powers developing and employing these weapons are not willing to reveal the inner workings of their software and hardware. Therefore, individual weapon systems should be examined and scrutinized based on whatever information is available in order to examine their levels of autonomy and meaningful human control.

It can be assumed that most people would have an adverse knee-jerk reaction when presented with the idea of intelligent machines making life-and-death decisions on the battlefield, and while those reactions may, to some extent, be based on fictional representations of intelligent killer robots, clearly the potential humanitarian harm that could result from the development and deployment of autonomous weapon systems presented in the analysed literature is a real danger, even if some of the highlighted issues are not unique to weapons with autonomous capabilities. Most of the arguments against autonomous weapon systems can, in one way or another, be traced back to the moral issue of not allowing machines to make life and death decisions about humans. This argument is not unique to autonomous weapon systems either and can be applied to other scenarios where artificial intelligence is put in the role of a critical decision maker, such as when self-driving cars must decide who to protect in traffic accident situations. However, autonomous weapon systems are perhaps the most straightforward example of this, and as previously mentioned, an examination of media coverage from recent years (Russell et al., 2021; Kallenborn, 2022) reveals that their advent is close if not already here.

The most intuitively appealing arguments that support the idea of humanitarian benefits of autonomous weapon systems rely on the fact that robots operating in place of human soldiers may lead to fewer casualties in war. While a futuristic scenario where all warfighting is conducted between autonomous weapon systems may seem intriguing, realistically humans will always be directly involved in conflicts, at the very least as civilians getting in the way. Be that as it may, Arkin's (2013) argument about autonomous weapon systems being potentially more capable of implementing a policy of not firing first due to having no inherent drive for self-preservation is worth considering. If the military powers developing autonomous weapon systems would decide to focus on implementing such features, it might indeed be possible that using autonomous weapon systems in place of human soldiers in some scenarios might lead to less combatant casualties on both sides and fewer accidental civilian deaths.

The issue of reliability was previously mentioned as just one of the many potential problems caused by the development and deployment of autonomous weapon systems, but many of the other issues are, in fact, derived from it. Even if autonomous weapon systems could behave more "humanely" on the battlefield than soldiers in some scenarios, it is doubtful whether this could realistically be true in all or even most cases. It is common

knowledge that software tends to have bugs, even more so when the software in question is complex. While previous research into testability of autonomous systems indicates that finding suitable methodologies for testing them is possible (Song et al., 2021), questions remain about whether the suggested methodologies could be utilized thoroughly enough to make sure that systems that directly make lethal decisions behave reliably and predictably in complex scenarios (Wallach, 2017; Righetti et al., 2018) that might be wildly different from the scenarios the systems were tested in. Autonomous weapon systems utilizing machine learning in their targeting decisions may improve their ability to distinguish between civilian and military targets over time, but their unpredictability may simultaneously increase (Wallach, 2017), and the idea that civilians might be unnecessarily killed due to an algorithm still being in the middle of a 'learning process' is unpalatable from the perspective of the international humanitarian law's principles of distinction and precaution.

While this thesis, as previously mentioned, accepts the requirements for meaningful human control as defined by Amoroso and Tamburrini (2020), the future of autonomous weapon systems will be defined by what level of meaningful human control will be required of them. Based on the analysed literature, it can be surmised that there is some level of agreement between political, academic and military actors regarding the importance of meaningful human control as a requirement for autonomous weapon systems. The discussion on meaningful human control boils down to the question of what level of autonomy is acceptable for lethal decision-making processes.

Some of the potential ways that autonomous weapon systems could be better than human soldiers at adhering to international humanitarian law advocated by Arkin (2013) are seemingly not achievable if meaningful human control is implemented into these systems. If a human operator is ultimately the one authorising the use of lethal force, the issues of possibly judgement-clouding emotions and the human limitations of situational awareness come back into play. Similarly, these systems would also have to operate outside of human accountability if they were to be making more humane decisions than soldiers. Because of the previously presented extensive arguments against unrestricted deployment and development of autonomous weapon systems, it is not reasonable to abandon the requirement for meaningful human control on the grounds of losing these purported humanitarian advantages. Arkin's (2013) arguments seem more relevant when considered in the context of a speculative future where AI technology has developed way beyond its current confines, while the call for meaningful human control (Wallach, 2017; De Sio & van den Hoven) is an immediate response to an emerging ethical issue.

From the analysed literature, value-sensitive design stands out as a potential methodology for ensuring that autonomous weapon systems stay under meaningful human control. If the welfare of civilians operating in conflict zones, who are indirect stakeholders as people whose lives are affected by autonomous weapon systems, are considered during the conceptual investigation stage of the value sensitive design process, and the guiding principles of distinction, proportionality and precaution of the international humanitarian law are implemented as values into the design process, it may be possible to design these systems to stay under meaningful human control. It is notable meaningful human control over autonomous weapon systems does not ensure their ethical behaviour on the battlefield; it simply maintains the status quo where humans are responsible for lethal decision-making. However, if the value of accountability is properly implemented into these systems, it may be possible to more easily trace who authorised certain actions as opposed to actions performed with conventional weapons.

6. Conclusions

The purpose of this thesis was to explore the following research questions through literature review:

What are the consequences of giving AI-based weapon systems the capability to make life and death decisions?

Could these systems be designed, developed and tested thoroughly enough so that their use in place of human soldiers would either mitigate the atrocities of war or at least cause no increase in them?

Because the current existence of autonomous weapon systems capable of independently performing deliberative life and death decisions is limited and debatable, the literature review had to rely in part upon academic speculation. Regardless, it was established that autonomous weapon systems are an emerging and contemporary real-world phenomenon that is currently being debated in academic literature. A working definition for what constitutes an autonomous weapon system was found by focusing on the capability for an independent, AI-based and deliberative process of target identification and engagement.

Criticism of autonomous weapon systems was the most common position taken by academics and organizations in the reviewed literature. The analysis of the reviewed literature revealed that the moral stance for not allowing autonomous weapon systems make life and death decisions over humans is backed up by substantial arguments regarding the possible negative humanitarian consequences that could stem from the issues of predictability, testability, reliability, vulnerability and accountability. The arguments were often presented in the context of a call for a complete ban on the development of autonomous weapon systems but were also found to be applicable as guidelines that could help with designing these systems to stay under meaningful human control.

Based on the reviewed literature, it was found that the proposed positive humanitarian consequences of autonomous weapon systems were for the most part based on a futuristic view that some may characterise as utopian speculation. The scenario where warfare becomes more humane as the result of humans handing over the reins of lethal decision making to intelligent machines seems difficult to achieve when the contemporary issues of autonomous weapon systems are compared against it, not to mention the fact that ethics may not be the first consideration of military and political stakeholders spearheading the development of these weapon systems. However, there are some situations where autonomous weapon systems might be able to operate in ways that reduce unnecessary casualties if they have been programmed to follow humanitarian principles such as ‘first-do-no-harm.’

As a part of an inquiry into the issue of testability of autonomous weapon systems, it was found that the current testing paradigms utilized for non-autonomous software cannot be directly applied to autonomous systems. Therefore, further research should be conducted in order to develop testing methodologies for ensuring the reliability and functionality of autonomous systems in general before developing ethically compliant autonomous weapon systems as described in the second research question becomes a realistic possibility.

Meaningful human control was found to be the most prominent of the suggested solutions to the issue of autonomous weapon systems, and therefore the most realistic direction for restricting the development of autonomous weapon systems in case a complete ban is not feasible, though further research is required in order to define concrete thresholds of human control that autonomous weapon systems must meet for the control to qualify as meaningful. Bringing human values to the forefront of design via the methodology of value sensitive design was found to be a potential approach for keeping autonomous weapon systems under meaningful human control. Because implementing meaningful human control into autonomous weapon systems may also hinder their strategic effectiveness, the motivation for military powers to focus on it may be limited. Therefore, academics and organizations should keep on highlighting the necessity of meaningful human control in order to draw more public attention to its importance.

A question of validity regarding the findings of this thesis may arise from the fact that a limited number of studies promoting the possible positive humanitarian effects of autonomous weapon systems were included as opposed to the number of those which offered critical perspectives on the phenomenon. This was due to the fact that there are not many prominent studies offering non-critical perspectives on the subject. As autonomous weapon systems and weapons technology in general is a heavy topic, it is conceivable that both this author and the cited authors are subject to some level of bias on the topic. Nonetheless, an effort was made to include different perspectives on the issue in the literature review. As real-world data on autonomous weapon systems and their consequences is currently lacking, this thesis should not be taken as an empirical evaluation, but as a collection of observations on the current state of academic debate and discussion on the subject.

Overall, this literature review indicates that according to the current state of research, the development and deployment of autonomous weapon systems has serious consequences from a humanitarian perspective. The efforts to design, develop and test autonomous weapon systems rigorously enough to mitigate their potential negative impacts should be focused on keeping them under meaningful human control when it comes to lethal decision making, even if the possibility that autonomous weapons could make more humane decisions than human soldiers in some situations is lost in the process. Because of the nature of autonomous weapon systems as weapons technology, performing quantitative research on the subject matter is difficult if not impossible without special access to military resources. However, findings from research into the development, testability and human control of autonomous systems in general could, in many cases, be applicable to autonomous weapon systems as well.

References

Alwardt, C. & Kruger, M. (2016). Autonomy of Weapon Systems. *Food for Thought Paper.* Institute for Peace Research and Security Policy at the University of Hamburg. Retrieved June 15, 2022, from https://ifsh.de/file-IFAR/pdf_english/IFAR_FFT_1_final.pdf

Amoroso, D. & Guglielmo, T. Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues (2020). *Current Robotics Reports,* 1, 187-194. Retrieved June 15, 2022, from https://link.springer.com/article/10.1007/s43154-020-00024-3

Arkin, R. (2008). Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture Part I: Motivation and Philosophy. In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

Arkin, R. (2013). Lethal Autonomous Systems and the Plight of the Non-combatant. *AISB Quarterly*, 137.

Arkin., R., Kaelbling, L., Russell, S., Sadigh, D., Scharre, P., Selman, B., Walsh, T. (2019). A Path Towards Reasonable Autonomous Weapons Regulation. *IEEE Spectrum.* https://spectrum.ieee.org/a-path-towards-reasonable-autonomous-weapons-regulation

Bouchet-Saulnier, F. (2013). *Practical Guide to Humanitarian Law*. Médecins Sans Frontíeres. Retrieved June 15, 2022, from https://guide-humanitarian-law.org/content/index/

Cellan-Jones, R. (2014, December 2). *Stephen Hawking warns artificial intelligence could end mankind.* BBC News. Retrieved June 15, 2022, from https://www.bbc.com/news/technology-30290540

Charters, D. (2020). Killing on Instinct: A Defense of Autonomous Weapon Systems for Offensive Combat. *Viterbi Conversations in Ethics*, 4(1). Retrieved June 15, 2022, from https://vce.usc.edu/volume-4-issue-1/killing-on-instinct-a-defense-of-autonomous-weapon-systems-for-offensive-combat/

De Vynck, G. (2021, July 7). *The U.S. says humans will always be in control of AI weapons. But the age of autonomous war is already here.* The Washington Post. Retrieved June 15, 2022, from https://www.washingtonpost.com/technology/2021/07/07/ai-weapons-us-military/

Dunlap, C.J. (2016). Accountability and Autonomous Weapons: Much Ado About Nothing? *Temple International & Comparative Law Journal*. 63-76. Retrieved June 15, 2022, from https://scholarship.law.duke.edu/faculty_scholarship/3592/

Felt, C. (2020, February 14). *Autonomous Weaponry: Are Killer Robots in Our Future?* The Henry M. Jackson School of International Studies, University of Washington. Retrieved June 15, 2022, from https://jsis.washington.edu/news/autonomous-weaponry-are-killer-robots-in-our-future/

Hernandez, J. (2021, June 1). *A Military Drone with A Mind Of Its Own Was Used In Combat, U.N. Says*. NPR. Retrieved June 15, 2022, from https://www.npr.org/2021/06/01/1002196245/a-u-n-report-suggests-libya-saw-the-first-battlefield-killing-by-an-autonomous-d

Holland Michel, A. (2021). *Known Unknowns: Data Issues and Military Autonomous Systems*. Geneva: United Nations Institute for Disarmament Research. Retrieved June 15, 2022, from https://unidir.org/sites/default/files/2021-05/Holland_KnownUnknowns_20210517_0.pdf

Caron, J.F. (2020). Defining semi-autonomous, automated and autonomous weapon systems in order to understand their ethical challenges. *Digital War*, 1, 173-177. Retrieved June 15, 2022, from https://www.researchgate.net/publication/346537479_Defining_semi-autonomous_automated_and_autonomous_weapon_systems_in_order_to_understand_their_ethical_challenges

Human Rights Watch. (2015). *Mind the Gap – The Lack of Accountability for Killer Robots.* HRW. Retrieved June 15, 2022, from https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots

Hynek, N. & Solovyeva, A. (2021). Operations of power in autonomous weapon systems: ethical conditions and socio-political prospects. *AI & Soc,* 36, 79–99.

Kallenborn, Z. (2021, October 5). *Applying arms-control frameworks to autonomous weapons.* Brookings Tech Stream. Retrieved June 15, 2022, from https://www.brookings.edu/techstream/applying-arms-control-frameworks-to-autonomous-weapons/

Kallenborn, Z. (2022, March 15). *Russia may have used a killer robot in Ukraine. Now what?* Bulletin of the Atomic Scientists. Retrieved June 15, 2022, from https://thebulletin.org/2022/03/russia-may-have-used-a-killer-robot-in-ukraine-now-what/

Klincewicz, M. (2015). Autonomous Weapon Systems: the frame problem and computer security. *Journal of Military Ethics,* 14(2), 162-176. Retrieved June 15, 2022, from https://www.researchgate.net/publication/281285922_Autonomous_Weapons_Systems_the_Frame_Problem_and_Computer_Security

Knopf, J.W. (2006). Doing a Literature Review. *Political Science & Politics*, 39(1), 126-132. Retrieved June 15, 2022, from https://core.ac.uk/download/pdf/81222467.pdf

McCoy, L., Burkell, J., Card, D., Davis, B., Gichoya, J., Le Page, S., Madras, D. (2019). *On Meaningful Human Control In High-Stakes Machine-Human Partnerships.* 2019 Summer Institute on AI and Society, AI Pulse. Retrieved June 15, 2022, from https://aipulse.org/on-meaningful-human-control-in-high-stakes-machine-human-partnerships/

Righetti, L., Pham, Q.-C., Madhavan, R. & Chatila, R. (2018). Lethal Autonomous Weapon Systems [Ethical, Legal and Societal Issues]. *IEEE Robotics*. 25(1). Retrieved June 15, 2022, from https://ieeexplore.ieee.org/document/8314620

Piper, K. (2019). *Death by algorithm: the age of killer robots is closer than you think.* Vox. Retrieved June 15, 2022, from https://www.vox.com/2019/6/21/18691459/killer-robots-lethal-autonomous-weapons-ai-war

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

Song, Q., Engström, E. & Runeson, P. (2021). Concepts in Testing of Autonomous Systems: Academic Literature and Industry Practice. *2021 IEEE/ACM Workshop on AI Engineering – Software Engineering for AI (WAIN).* 74-81. Retrieved June 15, 2022, from https://ieeexplore.ieee.org/abstract/document/9474374

Roff, H. M., & Moyes, R. (2016). Meaningful human control, artificial intelligence and autonomous weapons. In *Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems*, UN Convention on Certain Conventional Weapons.

Russell, S., Aguirre, S., Javorsky, E., Tegmark, M. (2021, June 16). Lethal Autonomous Weapons Exist; They Must Be Banned. *IEEE Spectrum*. Retrieved June 15, 2022, from https://spectrum.ieee.org/lethal-autonomous-weapons-exist-they-must-be-banned

Santoni De Sio, F., & van den Hoven, J. (2018). Meaningful Human Control Over Autonomous Systems: A Philosophical Account. *Frontiers In Robotics and AI*, *5*(15).

Wallach, W. (2017). Toward a ban on lethal autonomous weapons: surmounting the obstacles. *Communications of the ACM*, 60(5), 28–34.

Wyatt, A. (2020). So Just What Is a Killer Robot?: Detailing the Ongoing Debate around Defining Lethal Autonomous Weapon Systems. *Wild Blue Yonder Online Journal.* Retrieved June 15, 2022, from https://www.airuniversity.af.edu/Wild-Blue-Yonder/Article-Display/Article/2208774/so-just-what-is-a-killer-robot-detailing-the-ongoing-debate-around-defining-let/

Wyatt, A. & Galliott, J. (2021). An Empirical Examination of the Impact of Cross-Cultural Perspectives on Value Sensitive Design for Autonomous Systems. *Information 2021***,** 12(12), 527. Retrieved June 15, 2022, from https://www.mdpi.com/2078-2489/12/12/527

Zador, A. & Lecun, Y. (2019, September 26). Don't Fear the Terminator. *Scientific American Observations*. Retrieved June 15, 2022, from https://blogs.scientificamerican.com/observations/dont-fear-the-terminator/