Beyond the Ban: a comprehensive approach to the governance of autonomous weapons

Jia Yuan Loke
jyl37@alumni.cam.ac.uk
April 2018

Submitted in partial fulfilment of the requirements for the BA (Hons) in Human, Social and Political Sciences at the University of Cambridge

Abstract

Current debates about autonomous weapons tend to see a establishing a ban or developing regulations as mutually exclusive options. This dissertation argues that effective governance must incorporate both mechanisms. A legally binding treaty will most likely rest on a narrow, precise definition, meaning that it should be accompanied by forward- and backward-looking responsibility practices that go across states, the private sector and individual researchers. Borrowing from the field of nuclear politics, the dissertation further explores the pathways to creating an autonomous weapons taboo.

Table of Contents

Acknowledgements

From Cambridge POLIS, I thank Finbarr Livesey and Mette Eilstrup-Sangiovanni for their supervision.

Outside of my faculty, I am grateful to Shahar Avin and Haydn Belfield from CSER and Seth Baum from GCRI for their guidance. Special thanks to David Janku for helping me to find this topic.

Finally, I owe debt of gratitude to the GC4 community for all the tea and wraps.

At the Einsatzgruppen stage, the rounded-up victims were brought in front of machine guns and killed at point-blank range. Though efforts were made to keep the weapons at the longest possible distance from the ditches into which the murdered were to fall, it was exceedingly difficult for the shooters to overlook the connection between shooting and killing. This is why the administrators of genocide found the method primitive and inefficient, as well as dangerous to the morale of the perpetrators. Other murder techniques were therefore sought — such as would optically separate the killers from their victims. The search was successful, and led to the invention of first the mobile, then the stationary gas chambers; the latter — the most perfect the Nazis had time to invent — reduced the role of the killer to that of the 'sanitation officer' asked to empty a sackful of 'disinfecting chemicals' through an aperture in the roof of a building the interior of which he was not prompted to visit.
(Bauman, 2001, p. 26)

Introduction

This dissertation seeks to identify the best approach to governing autonomous weapons. Governance refers to a dynamic process that productively engages with "public and private actors, the activities of which are coordinated through both formal and informal rules and guidelines in such a way that a common or public goal is advanced" (Sending & Neumann, 2006, p. 653). In this case the goal is to mitigate the risks generated by autonomous weapons. This dissertation recognises that artificial intelligence *will* be incorporated into military systems, the question is *how.*

The dissertation proceeds in the following manner. First, an introduction to autonomous weapons and the current state of international discussions. Then the proposal of a heuristic framework to measure the difficulty of governing a class of weapons. Due to their high military utility, the fact that AI is a dual-use technology, and the moderate difficulties of establishing deployment, autonomous weapons are more challenging to govern than other prohibited weapons. The dissertation proceeds to identify six risks created by autonomous weapons. Although concerns about civilian casualties in direct combat have dominated discussions thus far, this dissertation suggests that the risks of escalating arms races or destabilizing wider military systems warrant greater attention. Given the challenges they pose and the risks they generate, this dissertation argues that effective governance of autonomous weapons is best achieved through three interacting mechanisms:

1. A legally binding ban on lethal autonomous weapons systems (LAWS)
2. Responsibility practices which guide the development of autonomous weapons
3. A taboo-like prohibition on the development and use of autonomous weapons

While the current discourse tends to see establishing a ban and developing regulations as mutually exclusive options, this dissertation suggests that effective governance should incorporate both approaches. The successful negotiation of a ban will most likely rest on a narrow, precise definition, which will not encompass all potential risks. Therefore, this ban should be accompanied by a set of responsibility practices which go across states, the commercial sector, and individual technical researchers to guide the development of machine autonomy. Finally, the first two mechanisms should be underpinned by an autonomous weapons taboo to ensure the highest likelihood of success.

Introduction to autonomous weapons

The term 'autonomous weapon' refers to an autonomous weapon system, which consists of three elements: a sensor, decision-making element and munition (Scharre, 2018). The system may be contained in a single unit or distributed across multiple platforms. Broadly speaking, autonomous weapons can be divided into three categories (Docherty, 2012):

Human-*in*-the-loop weapons search for targets, but the human operator decides which targets to engage. These are semiautonomous weapons.

Human-*on*-the-loop weapons detect and select targets and deliver force under the real-time oversight of a human operator. These are supervised autonomous weapons. Many defensive systems which seek to intercept incoming munitions fall into this category.

Human-*out*-of-the-loop weapons can search for targets, choose to engage, and deliver force without human input. Several terms have been used to describe this category of weapons, including fully autonomous weapons (FAWs); since 2013 states at the Convention on Certain Conventional Weapons (CCW) have settled on 'lethal autonomous weapons systems' (LAWS).

However, as addressed throughout this dissertation, the boundaries between these categories are not always clear.

The current state of governance

The current forum for discussing autonomous weapons is the United Nations Convention on Certain Conventional Weapons (CCW). Established in 1980, the CCW is an umbrella agreement between 125 states that provides a framework for prohibiting or restricting the use of certain weapons (United Nations Office for Disarmament Affairs, 2014). In November 2013, these states agreed to open talks on autonomous weapons; the first informal meeting occurred in May 2014 and a further two informal meetings were held in 2015 and 2016. In December 2016, the CCW formalized the discussion on autonomous weapons by establishing a Group of Governmental Experts (GGE); however the 2017 meeting of this newly established group was cancelled due to lack of funds. The annual CCW meetings resumed on 9-13 April 2018. At this meeting, the list of states calling for a ban on autonomous weapons grew to 26, while five countries (France, Israel, Russia, UK and the United States) explicitly rejected moving to negotiate new international law on autonomous weapons (Campaign to Stop Killer Robots, 2018).

The disarmament campaign against autonomous weapons became prominent in 2012 when Human Rights Watch (HRW) published a report calling for a preemptive ban on systems that have the ability to "select and engage targets without human intervention" (Docherty, 2012). In April 2013, a group of NGOs (including HRW) launched Campaign to Stop Killer Robots, "an international coalition working to ban fully autonomous weapons" (Campaign to Stop Killer Robots, 2018). As of April 2018, the Campaign consists of NGOs in 28 countries and remains the primacy civil society actor in the disarmament movement (Campaign to Stop Killer Robots, 2018).

Difficulties of governance

This dissertation proposes a heuristic to estimate the difficulty of governing a class of weapons. It argues that autonomous weapons are more difficult to govern than other prohibited weapons.

	Autonomous weapons	*Blinding lasers*	*Landmines*	*Chemical and Biological weapons*	*Nuclear weapons*
Military value	High	Low	Moderate	Moderate	High
Dual-use	High	Moderate-high	Moderate	High	Low
Difficulty of establishing deployment	Moderate	Low	Low	Low	Low

Military utility

Realism emphasises the role of material power in explaining political outcomes. In line with realist thinking, the biggest challenge to governance is the military utility of autonomous weapons. Scharre observes that "all military innovation since the first time a person threw a rock in anger has been about striking the enemy without putting oneself at risk" (Scharre, 2018, p. 276). Indeed, the chairman of the US Joint Chiefs of Staff has repeatedly stated that US service members should never have to face a fair fight (Gramone, 2016). Military leaders seek to conserve resources while increasing precision; therefore technologies that facilitate this are coveted (Corn, 2016). Autonomous weapons have been described as the 'third revolution in warfare' after gunpowder and nuclear weapons (FLI, 2015). For example, the US has turned to technology to compensate for a smaller military as part of its third offset strategy. During the first offset in the 1950s, Eisenhower stressed the importance of nuclear weapons as a deterrence to larger Warsaw Pact enemies. In the present day, the Department of Defence (DoD) has emphasised 'centaur' systems that integrate humans and computers. In March 2018, for instance, DARPA called for proposals as part of its OFFSET (Offensive-Swarm-Enabled Tactics) Program. By "leveraging and combining emerging technologies in swarm autonomy and human-swarm teaming", OFFSET envisions soldiers using small unmanned drones in swarms of 250 or more to accomplish missions (DARPA, 2018). Further examples which demonstrate the military utility of autonomous weapons are found in the 'problems of definition' section of this dissertation. It is worth nothing that while states continue to develop weapons with out-of-the-loop capabilities, most have indicated a desire – for now – to keep lethal targeting decisions under human control.

Dual-use

"Any AI research could be co-opted into the services of war, from autonomous cars to smarter-chat bots…it's a short hop from innocent research to weaponization" (Dvorsky, 2015). The extent to which a weapon (and the technology that enables it) is dual-use affects how available said technology is outside of the military domain, which in turns informs the scope of regulatory efforts. Borrowing from the field of life sciences, dual-use technology refers to "research that is intended for benefit, but which might easily be misapplied to do harm" (WHO, 2018). For instance, open-source facial recognition, navigation and planning, and multi-agent swarming algorithms are developed for beneficial use but are equally applicable to the development of autonomous weapons. Moreover the field of AI emphasises the sharing of ideas and research.

A great deal of AI development is located in commercial rather than military settings. Cummings observes that "military autonomous systems development has been slow and incremental at best, and pales in comparison with the advances made in commercial autonomous systems" (Cummings M. L., 2017, p. 1). In 2016, the aerospace and defence industries spent US$30 billion on research and development to enable autonomy, while the automotive sector spent US$95 billion and the information & communication sector spent US$200 billion (Cummings M. L., 2017).

Unlike the development of nuclear weapons, a large portion of the third offset will rely on external contractors. Between 2015—2017 Secretary of Defence Ashton B Carter visited Silicon Valley multiple times; before that it had been 20 years since a defence secretary visited that area (Markoff, 2016). In March 2018, it was reported that Google is partnering with the US DoD to develop algorithms to analyse drone footage as part of the Algorithmic Warfare Cross-Functional Team (Project Maven), which seeks to "accelerate the DoD's integration of big data and machine learning" (Deputy Secretary of Defense, 2018). In April 2018 the DoD announced the creation of the Joint Artificial Intelligence Centre (JAIC); a report from the Defence Innovation Board states that the JAIC should "liaise with other labs in the private sector and universities" (Defense Innovation Board, 2018). This board is chaired by former Google chief executive Eric Schmidt.

Establishing deployment

This refers to how distinguishable a weapon is from conventional weapons. The rationalist perspective suggests that actors seek to misrepresent their true capabilities in order to obtain a military or strategic advantage (Fearon, 1995). The less detectable a weapon is, the more likely it can be deployed without compromising that advantage. Establishing deployment is a function of distance and visibility; underwater and aerial environments are particularly opaque. In April 2018, an Israeli-made Harop missile-drone destroyed a bus in the disputed area of Nagorno-Karabakh, killing seven. The Harop has the capability to be fully autonomous, and it is difficult to establish if there was a human operator behind this attack. In 2011, the UK deployed 24 'fire-and-forget' Brimstone missiles, destroying eight or more Libyan tanks simultaneously. "It would have been difficult", suggests John Markoff, "for human operators to coordinate the swarm of missiles with similar precision" (Markoff, 2016). As Mary Wareham, coordinator of the Campaign to Stop Killer Robots, has conveyed, "there is a constant debate inside the Campaign as to when we remove the word 'preemptive' from our call for a ban" (Perrigo, 2018).

Risks

Having established that autonomous weapons present a formidable challenge to governance, we must next ask ourselves what sort of governance is necessary. This in turn demands analysis of the risks we seek to mitigate. This dissertation identifies six risk areas associated with autonomous weapons. Discussions thus far has been framed by the humanitarian disarmament campaign, therefore concerns about civilian casualties and proliferation to rogue actors have been most prominent. Meanwhile "there has been virtually no discussion of how autonomous weapons might complicate crisis stability [or] escalation control" (Scharre, 2018, p. 351). While recognizing that the current approach may be necessary to engender public support, the dissertation suggests that other risks warrant further attention.

Casualties in direct combat

The most visible risk created by autonomous weapons arises from their inability to distinguish between lawful and unlawful targets, assess the proportionality of attack, and avoid causing unnecessary suffering as required by International Humanitarian Law (IHL). The primary concern of the 2012 HRW report was "the impact of FAWs on the protection of civilians during times of war" (Docherty, 2012). In 2013 the International Committee for Robot Arms Control (ICRAC) issued a statement endorsed by 272 experts from 37 countries: "in the absence of clear scientific evidence that robot weapons have, or are likely to have in the foreseeable future, the functionality required for accurate target identification, situational awareness or decisions regarding the proportional use of force, we question whether they could meet the strict requirements for the use of force" (ICRAC, 2013). The ability of autonomous weapons to perform in cluttered open-textured environments remains uncertain.

However, others suggest that machines may outperform humans in combat. While recognising that the necessary technology does not exist today, Arkin argues that there is no "fundamental scientific limitation to achieving the goal of these machines being able to discriminate better than human can in the fog of war" (Arkin, 2009). Machines do not have self-preservation as a foremost drive, allowing them to act conservatively. They can be designed without emotions such as fear and anger that cloud judgement (Arkin, 2007). At the 2018 CCW meeting, the US and UK opposed a preemptive ban on the grounds that machine autonomy has the potential to enhance civilian protection (Evans, 2018). Charles Dunlap argues that banning a class of weapons based on a "technological 'snapshot of time'" denies the possibility for later advancements to reduce casualties (Dunlap, 2015). For example, technological advances have allowed precision-guided munitions to replace the indiscriminate bombing campaigns of WWII (Scharre, 2018). From this perspective, "banning is like Luddism", and a ban may prohibit research on potentially beneficial uses of autonomy (Arkin, 2018, p. 284).

In response, ban advocates counter that even if future weapons are capable of adhering to IHL, they may pass through a problematic 'teething' phase, a dangerous middle ground of poor judgement (Roff, 2016). Although states are disincentivised from fielding unreliable autonomous weapons because doing so may harm their own soldiers, the fact that warfare environments are increasingly becoming asymmetric means that risks of dangerous behavior are shifting to foreign civilians. Therefore states may deploy a 'test as you go' to enable faster weapons development (Bhuta & Pantazopoulos, 2016). As will be discussed later, actors may further be encouraged to field unreliable autonomous systems because they may not be held accountable if these systems make a mistake.

System instability

The proliferation of machine autonomy across military systems may lead to the telescoping of timescales in conflict. This is particularly relevant for defensive systems and in cyber warfare. Carvin applies Perrow's normal accident theory to argue that machine autonomy increases the incomprehensibility of complex systems, raising the likelihood of accidents in high-risk systems. Further, it will be challenging to investigate why an accident has occurred and to assign responsibility for it (Carvin, 2017). Machines make decisions far quicker than the human mind, pushing adversaries to act and react fast. "Sun glint in visual data misinterpreted as a rocket flame, sudden, unforeseen moves of the enemy swarm, a simple software bug could trigger an erroneous counter-attack" (Altmann & Sauer, 2016). While the error may start at a small scale, the sequence of events developing from two (or more) autonomous systems interacting at high speed may lead to the escalation of otherwise controllable crises. The effects of rapid algorithmic interactions are demonstrated in the May 2010 stock market 'flash crash', during which the Dow Jones fell by 573 points in five minutes as autonomous systems interacted in an uncontrolled feedback loop (Salmon & Stokes, 2010). During the 1983 nuclear near-miss, Soviet radars incorrectly reported an American attack (a false positive). Despite the context of severe tension, Stanislav Petrov had the *feeling* it was a computer error and dismissed the warning as a false alarm, plausibly preventing a Soviet nuclear counter-attack. In the present day, Lora Saalman observes that the Chinese military has a fear of false negatives in its alarm systems (a strike occurs but the system does not identify it). To the extent that China integrates AI into its military, it will likely seek to address this perceived deficiency in its early warning systems (Saalman, 2018).

Proliferation to rogue actors

The aforementioned examples have largely assumed that actors pay attention to IHL; this is not always the case. The Future of Life Institute (FLI) postulates that "if any major military power pushes ahead with AI weapon development…autonomous weapons will become the Kalashnikovs of tomorrow" (FLI, 2015). Due to the dual-use nature of machine autonomy, the barriers to developing autonomous weapons are lowered. The concern is not so much sophisticated LAWS but 'DIY' autonomous weapons; commercial drone sales in the US rose from US$44M in 2013 to US$1.3B in 2017 (Dunn, 2017). In April 2018, footage emerged allegedly showing pro-Ukrainian actors using commercial drones to drop munitions on pro-Russian actors near Mariupol, Ukraine.[1] FLI goes on to suggest that "it will only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control the populace, warlords wishing to perpetrate ethnic cleaning" (FLI, 2015). In response Arkin's arguments, Carpenter points out while robots may be designed without rage, they may equally be designed without empathy, a conscience, or other emotions prevent human soldiers from carrying out the most abhorrent of atrocities (Carpenter, 2014). Thus they may be the perfect tools of repression, torture or rape.

Arms race escalation

Viewed through the lens of any one military, autonomous weapons have immense utility. When two (or more) sides possess them, however, an arms race may be triggered. FLI states that the "the key question for humanity today is whether to start a global AI arms race or to prevent it from starting" (FLI, 2015). Similarly, the Campaign to Stop Killer Robots suggests that if states choose to develop autonomous weapons, "others may feel compelled to abandon policies of restraint, leading to a robotic arms race" (Campaign to Stop Killer Robots, 2013). Almost 100 states and non-state actors including Hamas, Hezbollah, and ISIS currently possess drones. (Scharre, 2018). As it stands, information about the state of development in Russian and Chinese systems remains limited (Roff, 2018). Arms races exacerbate errors of judgment, risk-taking, and hostile motivations (Armstrong, Bostrom, & Schulman, 2013). In 2017, Putin commented that "whoever becomes the leader in this sphere will become the ruler of the world".

[1] https://www.youtube.com/watch?v=tuLUFqaC02Q

Increased tension may be accompanied by lower costs of entering into conflict. At the 2018 CCW meetings, Pakistan argued that replacing human troops with machines "will lower the threshold of going to war; consequently, the resort to use of force will become a more frequent" (Permanent Representative of Pakistan, 2018, p. 1). Richard John Gatling, inventor of the machine gun, believed that his creation would save lives by reducing the number of soldiers needed on the battlefield; instead it allowed killing at an unpreceded rate.

The broader field of AI

Developing machine autonomy for military purposes may sabotage the wider goal of building fair, accountable, transparent and safe (FATS) AI. The military utility of AI may push actors to skimp on FATS precautions in order to gain superiority, resulting in a situation where teams 'race to the precipice' in an AI arms race (Armstrong, Bostrom, & Schulman, 2013). Allowing AI to become associated with autonomous weapons may tarnish the field, creating a public backlash against AI which hamstrings the further development of beneficial technology (FLI, 2015). The historical antagonism towards nuclear technology has arguably handicapped the development of nuclear power as an alternative to fossil fuels. Weinberg stated that "the public perception and acceptance of nuclear energy…has emerged as the most critical question concerning the future of nuclear energy" (Weinberg, 1975, p. 19). This dissertation believes that AI is likely to exceedingly beneficial for humanity in areas including healthcare, science and transportation; we should seek not to curtail these potential benefits.

Moral concerns

The deontological position argues that the decision to kill humans should not be transferred to a machine. From this perspective, allowing robots the power and authority to kill humans beings "crosses a fundamental moral line" (Campaign to Stop Killer Robots, 2013). Ban proponents argue that autonomous weapons violate the Martens Clause, which appears across international instruments including the 1977 Additional Protocol to the Geneva Conventions as well as the preamble to the CCW. The Martens Clause implies that weapons which contravene the 'principles of humanity' and the 'dictates of public conscience' should not be used even if no specific prohibition exists yet (Goose & Wareham, 2017). "Take out the human", argues Peter Lee, "and you also take out the humanity of the decision to kill" (Lee, 2018).

Little to gain, much to lose

In the current literature, much has been said about each of the various risks, but less effort has been made to compare between then. Most analysis go to the extent of arguing that "these risks…are so large that the weapons posing them must be banned outright" (Roff, 2016). The dissertation therefore attempts to weigh the magnitude of various risks. It argues that the potential benefits of autonomous weapons outperforming humans in direct combat are outweighed by other risk areas. These benefits, if they arise, will have a bearing on combat situations that involve technologically advanced IHL-compliant actors. However such situations represent a small portion of all combat. 2016, for example, saw 1.2 battle deaths per 100,000 or 90,000 total battle deaths (Pinker, 2018). Most of these deaths arose from 12 civil wars, involving actors that are unlikely to prioritise IHL. The most significant of these civil wars is the ongoing conflict in Syria which accounted for 50,000 of the 90,000 battle deaths; this conflict has seen numerous violations of IHL including multiple chemical attacks (Syrian Observatory for Human Rights, 2018). Assuming (generously) that 10% of total deaths in 2016 involved IHL-compliant actors, and that machine autonomy increases precision by 50%, expected gain is 4,500 lives. Other things equal, if autonomous weapons proliferate and increase the lethality of non-IHL compliant conflict by just 10%, the expected loss is 8,100 lives.

This dissertation argues that the most significant risks are those that increase the likelihood of great power conflict. Such wars are responsible for the majority of the victims of all wars combined. Consider a situation where a false positive by an early warning system leads to conflict between great powers. As a proxy, WWII saw 300 battle deaths per 100,000 or 7.5 million deaths per year (excluding deaths from disease and famine) (Pinker, 2018). Increasing the likelihood of such a conflict by 1% represents an expected loss of 75,000 lives. Even if precision reduces casualties by 50% the expected loss is still 37,500. The potential benefits of AI for wider society are tremendous – reducing that potential by even 1% represents significant loss of expected value. All this is before accounting for moral concerns.

This dissertation recognises that the above thought experiments make serious assumptions about complex dynamic systems. There remains a great deal of uncertainty about the capabilities of autonomous weapons as well as the behavior of states and non-state actors. However the point remains that the introduction of autonomous weapons affects high-consequence outcomes. The fact that battle deaths per 100,000 have declined from 300 in 1950 to 1.2 in 2016 indicates that when it comes to warfighting, we do not stand at a neutral point. There are gains to be made, but far more to lose. Therefore, in line with the precautionary principle, this dissertation supports a targeted legally binding ban on LAWS.

A legally binding treaty

Given the significant risks posed by autonomous weapons, this dissertation puts forward that a legally binding ban is desired. From a contractual perspective, such an agreement will help to solve problems of coordination created by the significant military utility of autonomous weapons. A clear ban strengthens the credibility of commitment, lowering arms race risks by removing ambiguity and reducing the space for slippage. It addresses the moral hazards created by autonomous weapons. A legally binding treaty also instantiates normative values (Abbott & Snidal, 2000). This dissertation proceeds to investigate one of the primary conceptual challenges to achieving said ban – establishing clear definitions of 'LAWS' and 'meaningful human control'. It argues that focusing on a narrow definition presents the highest likelihood of success.

Problems of definition

Any regulatory regime must define what exactly it is that the regime bans, therefore it must define LAWS. The issue of defining autonomous weapons – and the extent to which definitions are necessary – has been one of the main topics of discussion at CCW meetings (Vilmer, 2016). "Within and outside the field of computer science", observe Bhuta and Pantazopoulos, "no stable consensus exists concerning the meaning of autonomy in weapons systems." (Bhuta & Pantazopoulos, 2016, p. 284). At the 2018 meeting, states continued to debate the definition of terms including 'maintaining', 'ensuring', 'substantiate' and 'appropriate' (Campaign to Stop Killer Robots, 2018).

States are developing and deploying ever more sophisticated systems that blur the distinction between on-the-loop and out-of-the-loop. Arkin observes a trend toward "the application of lethality by autonomous systems" (Arkin, 2007). Kalmanovitz notes that military technology has been moving toward autonomy for decades, and that the full autonomisation of lethal weapons is just another step in an ongoing trend (Kalmanovitz, 2016). As of April 2018, a dataset compiled by Heather Roff lists 284 weapons systems with autonomous features developed or deployed by 18 states. Her findings show that the states with the most autonomous systems are the US, Russia and China. Target discrimination and self-engagement are the two most recent emerging technologies; they represent the current frontier of autonomous weapons. We can observe an 'autonomy creep' as systems begin to trend into the area where "the weapon does not have a specific target but a set of potential targets, and it waits in the engagement zone until an appropriate target is detected" (Roff, 2018). Examples of systems with the highest degrees of autonomy include:

Name	*Country*	*Operational date*	*Description*
AGM 185C LRASM (2018)	USA	2018	Long range anti-ship missile expected to be capable of autonomous targeting, evasion of active defence systems, and independent acquisition of the target without prior intelligence or support from GPS
TARES	USA	2020	UAV designed to be launched in groups of up to 16 with the ability autonomously navigate and loiter for up to 4 hours. Can select, prioritise and decide to engage targets while communicating with other TARES
nEUROn	France	Development	UAV with fully autonomous attack capabilities, target adjustment, and information sharing between systems
Taranis	UK	2013	UAV with the capability to land, pilot to area of interest, loiter, and attack autonomously; currently remotely piloted but has the capability to be autonomous

(Roff, Dataset: Survey of Autonomous Weapons Systems, 2018)

The inverse of full autonomy is maintaining meaningful human control, and here pro-regulation actors have struggled as well. Crootof notes that "to the extent that there is any consensus among states, ban advocates, and ban sceptics regarding the regulation of autonomous weapon systems (AWS), it is grounded in the idea that all weaponry should be subject to 'meaningful human control'" (Crootof, 2016, p. 53). As the Czech Republic has observed, "the decision to end somebody's life must remain under meaningful human control…the challenging part is to establish what precisely 'meaningful human control' would entail" (Statement by the Czech Republic, 2015, p. 1). The ICRAC has put forward three conditions of meaningful human control: commanders should have full awareness of the target area, cognitive participation in the attack, and the means to rapidly abort of the attack (Crootof, 2016). From this perspective, however, even catapults are unlawful – humans employing them have neither real-time awareness of the target area nor the ability to suspend the projectile after launch. Human on-the-loop weapons select a target for human approval; this type of control has been shown to create automation bias, which refers to a class of judgment errors where humans become overconfident about computer generated decisions and discount the need to search for contradictory information (Cummings M. , 2006). As we move along the spectrum from automated to fully autonomous, what is the tipping point where humans are no longer meaningfully in the loop?

Definitions as a site of contestation

We should consider that establishing definitions is a mode of contestation and negotiation between different actors. "The variety of conceptions of machine autonomy in the discourse of autonomous robots reflects the many ideas, ambitions and goals of the various social groups with a stake in the development of these technologies" (Noorman & Johnson, 2014, p. 56). At the 2016 CCW meetings, China suggested that a specific definition of autonomy was a prerequisite for further discussions, which others interpreted as a method blocking progress. Overall the question was less the definition of autonomous weapons but whether it was necessary to have a concrete definition; a consensus was reached that a working definition should be used to progress discussions (Vilmer, 2016).

The United Sates is the only country which has a clear written policy guiding the development of LAWS, establishing the concept of 'appropriate levels of human judgement' through Directive 3000.09 in 2012. Dan Saxon argues that "the Directive's neutral language and parameters for the design and use of AWS is intended to encourage the development of new autonomous technologies…in doing so Directive 3000.09 removes political-bureaucratic obstacles to the development of AWS"; it empowers "the US desire to maximise available options in the interests of military necessity" (Saxon, 2016, p. 198). This is indicative of the ability of the US to simultaneously break and define convention in global affairs.

Establishing definitions

This dissertation observes that developing effective definitions to govern autonomous weapons represents a trade-off between clarity and scope. The capabilities of machine autonomy present a rapidly moving target; therefore this dissertation suggests that focusing on the human role presents a good approach. Scharre observes that "it is freedom, not intelligence, that defines an autonomous weapon" (Scharre, 2018, p. 50). This is the method that ban advocates have begun to adopt. HRW "does not oppose military usage of autonomy and technologies related to artificial intelligence, but it draws the line at the development of machines that could select and fire on targets without human intervention" (Goose & Wareham, 2017). Professor Noel Sharkey, chair of the ICRAC, has remarked that "we are only interested in banning the critical functions of target selection and applying violent force, two functions" (Perrigo, 2018). Such an approach addresses the moral risks created by autonomous weapons. It should not hinder the development of potentially beneficial technology. As states continue to discusses what constitutes meaningful human control, the dissertation emphasises that policymakers must work closely with technical researchers to design and enshrine appropriate points of human interaction in complex military networks that make high-consequence decisions. Altogether, a precise definition provides a necessarily clear focal point for the legislative process, generating a higher likelihood that a legally binding arms control agreement will be successful.

A targeted definition of LAWS, however, does not comprehensively capture the risks generated by the progressive autonomations of weapons systems (Bhuta & Pantazopoulos, 2016). Saxon's analysis highlights that in clearly defining and banning an ideal-type, we may implicitly permit everything else up till that ideal. The military utility of machine autonomy means that states will continue to develop semiautonomous and supervised autonomous weapons. Weapons do not need to be fully autonomous in order for them to have harmful impacts. Furthermore, the likelihood and timeframe of achieving a legally binding ban remains uncertain. Adoption of said ban will involve significant contracting costs (Abbott & Snidal, 2000). Even by the standards of international discussions, the CCW process is known to be slow and prone to failure. Due to the consensus-based decision-making system, outcomes are decided by the lowest common denominator. Therefore, ongoing efforts to establish hard legislation should be accompanied by the establishment of responsibility practices that encompass the development and use of autonomous weapons.

Responsibility practices

The current discourse tends to see a ban and regulation as mutually exclusive options; HRW, for example, "has urged that states ban rather than regulate fully autonomous weapons" (Goose & Wareham, 2017). However this dissertation argues that a targeted ban should be supplemented by responsibility practices that go across a number of actors including states, the commercial sector, and individual technical researchers. The notion of responsibility practices refers to "established ways that people, within a particular environment or community, understand, evaluate, and ascribe responsibility". (Noorman M. , 2013, p. 813). They "involve accepted ways of evaluating actions, holding others to account, blaming or praising, and conveying expectations about obligations and duties" (Johnson & Noorman, 2014, p. 18). These practices may be "formal and informal, comprising law and legal devices…as well as organizational and professional norms" (Bhuta, Beck, & Geib, 2016). Responsibility practices involve complementary *ex ante* regulations and *ex post* liability components. By adopting responsibility practices, technical experts uphold and improve their reputation, ensuring that beneficial applications of AI are not curtailed.

Problems of responsibility

The lack of forward-looking responsibility practices may be attributed to the dual-use nature of machine autonomy as well as the uncertain timeline and capabilities of autonomous weapons development. Scherer observes that AI research is diffuse and discrete. Diffuse means that the individuals working on the components of AI systems may be located far away from one another; there are a large number of open-source machine learning libraries. Discreteness means that separate parts of an AI system can be designed without consciousness coordination. Any individual who contributes to an open-source library does not know beforehand what other it may be used for in the future (Scherer, 2016).

Furthermore, forward-looking practices are currently lacklustre because of the absence of pressure from backward-looking mechanisms. If individuals are held accountable for their actions, they may be less likely to commit them again. This will also discourage others from performing similar acts for fear of being punished. Shelton notes that "the deterrence literature shows a correlation between the certainty of consequences and reduction of offences" (Shelton, 2005). Accordingly, the promise of punishment should be common knowledge so that actors can weigh the anticipated costs of their actions. This dissertation proceeds to illustrate how autonomous weapons create challenges for the establishment of direct or indirect responsibility.

The doctrine of direct responsibility holds actors liable for playing an active role in the perpetration of a crime. However, it is difficult to assign direct responsibility to autonomous weapons for three reasons. First, actions committed by machines do not fulfil the *mens rea* element – the mental state of committing a crime. Second, international criminal tribunals "have jurisdiction over natural persons" rather than robots (International Criminal Court, 1998, p. 17). The Nuremberg Tribunal explained that "crimes against international law are committed by men, not by abstract entities, and only by punishing individuals who commit such crimes can the provisions of international law be enforced" (Judgement of the Nuremberg International Military Tribunal, 1947, p. 55). Third, it is difficult to punish autonomous weapons because they do not experience suffering the way humans do. As the would-be-perpetrators are replaced by autonomous entities, we may look to commanders as the responsible subjects.

Indirect responsibility is the legal doctrine that holds a commander liable for failing to take necessary and reasonable measures to prevent or punish the crimes of a subordinate (Docherty, 2015). Across various permutations in domestic and international law, the establishment of indirect responsibility is repeatedly hamstrung by the nature of machine autonomy. Indirect responsibility arises when a commander has knowledge of a crime, the ability to prevent crimes, or the ability to exert effective control over the subordinate (Docherty, 2015). LAWS, by definition, select and engage targets without human intervention. Jain suggests that that "the primary barrier for establishing accountability for the harm caused by an AWS is the epistemic uncertainty associated with its conduct, which is a deliberate part of its design features" (Jain, 2016). Much of the value of AI (in war and other contexts) depends on it not being predictable in certain situation nor so transparent that humans can understand it. We see this in AI systems that play GO, which executes moves that go against established doctrines of human strategy. As previously explored, machines make decisions in timescales far quicker than the human mind. Moreover, after these decisions are made, they are not easily reverse engineered, audited or explained. Deciphering the 'black box' problem remains a significant technical challenge for the machine learning community (FAT ML, 2016). Aeroplanes do not flap their wings; autonomous weapons do not 'think' as soldiers do. Overall we see that existing frameworks for generating responsibility are stretched by machine autonomy – therefore we should look to those who engineer it.

Establishing responsibility

Most ban advocates point to the responsibility gap as a reason to ban autonomous weapons; HRW suggests that "states should eliminate this accountability gap by adopting an international ban on fully autonomous weapons" (Docherty, 2015). However this dissertation suggests that closing the responsibility gap in the direction of weapons developers and technical researchers is a valuable way to accompany a narrow ban. As Johnson notes, in response to emerging technologies that create a responsibility gap, society has in the past "set up mechanisms to pressure those who make and use these technologies to operate safely and to take responsibility when something goes wrong". (Johnson D. G., 2015, p. 714).

If we conceive of the soldier as an autonomous unit, the technical development of autonomous weapons is equivalent to the training, guidance and discipline that soldiers undergo. Preparing soldiers to obey commands and subordinate their self-interest to the aims of the mission is the goal of military leadership. Customary international humanitarian law, including the 1949 Geneva Conventions, places an obligation on states to 'teach' their soldiers. From this perspective, "only those individuals capable of autonomous reasoning who have been incorporated into an organization capable of managing the exercise of that reasoning should be granted the privilege of engaging in hostilities" (Corn, 2016, p. 222). Field commanders, however, do not have the opportunity to influence the judgement of autonomous weapons through military training and rely almost exclusively on the 'training' provided by those who develop the weapons. Therefore the focal point of compliance should shift to the those who program and develop weapons.

The field of biotechnology provides a template for developing responsibility practices. Like machine autonomy, biotechnology is an area with a high degree of dual-use concerns. The Biological Weapons Convention (BWC) prohibits the hostile application of the life sciences. It entered into force in 1975, building on 1925 Geneva Protocol that bans chemical and biological weapons. As a disarmament treaty the BWC was initially focused on states but has expanded its focus to non-state actors; at the 5th review conference in 2001, states agreed to discuss a code of conduct of scientists (UNOG, 2018). We can further look to domestic regulations. In 2012, the United States began implementing Dual-Use Research of Concern policies; these policies establish a review system at the publication stage to raise concerns about potentially harmful knowledge (Office of Science Policy, 2012).

Bhuta and Pantazopoulos argue that international environmental law (IEL) provides a relevant set of principles for addressing *ex ante* uncertainty. To address this, IEL places special obligations of care and strict burdens of liability on states. Under this framework, "states engaged in an activity the consequences of which entail a risk of causing significant transboundary harm may be responsible even if the activity is not in itself unlawful" (Bhuta & Pantazopoulos, 2016, p. 291). Actors have a duty to identify, comprehend, and minimise risks that may harm others. The IEL framework rests on the idea of 'active vigilance', which implies an obligation to constantly stay abreast of technological development and to take appropriate measures to address risks. The IEL framework creates legal obligations of due diligence, risk assessment, information sharing, monitoring and technical standard setting for states who develop autonomous weapons (Bhuta & Pantazopoulos, 2016).

There are signs that the field of AI is beginning to take a wider view of its responsibilities in the military sphere. In March 2018, more than 50 academics from

30 countries signed a letter calling for a boycott of Korea Advanced Institute of Science and Technology (KAIST) after it was announced that KAIST would be collaborating with Hanwha, South Korea's leading arms company, to develop "AI technologies to be applied to military weapons, joining the global competition to develop autonomous arms" (Boycott of Kaist, 2018). The boycott was called off after KAIST committed not to develop LAWS. In April 2018, some 3,100 Google employees (out of more than 70,000) signed a letter protesting the company's involvement in Project Maven. Arguing that "Google should not be in the business of war", the letter called on Google to cancel involvement in the project and create a clear company policy stating that it will not build warfare technology (Letter to Google C.E.O, 2018, p. 1). The Institute of Electrical and Electronics Engineers (IEEE) is currently drafting an 'Ethically Aligned Design' document to guide the responsible development of AI. Included in the series of detailed recommendations is the recognition that "professional ethics…can and should have a higher standard covering a broader array of concerns" than legally binding international agreements on autonomous weapons (IEEE Global Initiative , 2016, p. 68). It reminds researchers that "the chains of accountability backward, and predictability forward, also include technical aspects" (IEEE Global Initiative , 2016, p. 68). This represents a major effort by the technical community to develop responsibility practices concerning autonomous weapons. The dissertation recommends the creation of a review system that brings together the research and policymaking communities to assess research at the publication stage. One can imagine how unreserved publishing of adversarial examples (inputs to machine learning models that are designed to cause the model to make a mistake) can generate security risks. Researchers should strive to consider the implications of their work for policy, and policymakers must seek to understand and engage with the technology at hand.

Thus we have examined the creation of forward-looking and backward-looking responsibility practices to promote a wider set of obligations on those who develop weapons. However establishing proximate cause for technical researchers will prove difficult, and codes of conduct may not be legally binding. The problem of governing autonomous weapons remains vulnerable to the 'unilateralist's curse' – for a positive outcome to occur, all actors must independently decide to act impartially and in good faith. A negative outcome, however, requires the unilateral move of one misguided or malicious actor. Outcomes are held hostage to the judgement of the most extreme outlier. Even if most states would rather not develop autonomous weapons, they may be forced to do so in the interest of national security. To improve the chances that governance will be successful, we can draw lessons from a domain where the standards of behavior have *not* been created by and for the powerful.

An autonomous weapons taboo

We have broached the idea of norms in the discussion on responsibility practices. According to Tannenwald, a taboo is "a particularly forceful kind of normative prohibition that is concerned with the protection of individuals and societies from behavior that is perceived to be dangerous" (Tannenwald, 2005, p. 8). Nuclear weapons present a case-study where a class of weapons have been restricted despite them possessing immense military utility. There has been little discussion in the literature about a taboo on autonomous weapons, and even less that explores the pathways to developing one. This dissertation postulates that one or both of the previous governing mechanisms (as well as the efforts to establish them) may instantiate the formation of a taboo-like prohibition on autonomous weapons. In turn, this taboo lends force to disarmament efforts by "[shaping] the realms of possibility…[influencing] the probability or occurrence of certain courses of action (Tannenwald, 2008, p. 4). It proceeds to examine how a taboo-like prohibition against nuclear weapons developed, analyse the extent to which similar conditions exist for autonomous weapons, and suggest ways to strengthen current efforts.

The nuclear taboo

Tannenwald argues that a powerful taboo against the use of nuclear weapons has developed in the global system, which, although not a fully robust prohibition, has stigmatized nuclear weapons as unacceptable (Tannenwald, 2005). Paul suggest that the tradition of non-use is not yet a taboo but rather an 'informal social norm' (Paul, 2009). To incorporate both ideas this dissertation uses the term 'taboo' in a loose sense to refer to unwritten mechanisms that operate alongside the material considerations of states. These mechanisms lead to revulsion towards the objects of taboo and create widely held inhibitions on their use. Although not codified in law, a taboo shapes and restrain the self-help behavior of states.

Today nuclear weapons are not viewed as 'just another weapon', but historical records suggest that this was not always the case. High-level American officials produced plans to use nuclear weapons in the 1940s and 1950s, such as during the Korean War. Thereafter, one serious threat of nuclear war occurred in the 1962 Cuban missile crisis, but overall little evidence exists that high-level officials have seriously considered using nuclear weapons. US leaders barely considered using nuclear weapons during the 1991 Gulf War. States such as Sweden and Switzerland once assumed that they would acquire nuclear weapons as part of the process of acquiring the latest weapons technology, but this is no longer the case (Tannenwald, 2005). While the tradition of non-use between nuclear states may be accounted for by the dynamic of mutually assured destruction, it is more puzzling to explain the non-use of nuclear weapons against non-nuclear states. Non-nuclear states have initiated conflict against nuclear states, anticipating the non-use of nuclear weapons by nuclear states. Nuclear states have continued fighting and incurred significant costs rather than resort to nuclear weapons to terminate the war swiftly. The Vietnam War presents a case where the United States chose to prolong a frustrating war rather than 'win' by using nuclear weapons (Paul, 2009).

All this is despite the fact that before 2017, international law did not explicitly prohibit nuclear weapons. In 1996 the International Court of Justice (ICJ) ruled that "the Court cannot conclude definitively whether the threat of use of nuclear weapons would be lawful or unlawful in an extreme circumstance of self-defence, in which the very survival of a State would be at stake" (ICJ, 1996, p. 44). Further, the nuclear taboo continues to develop in the face of long-standing official resistance by the US government and other nuclear powers (Tannenwald, 2005). No nuclear states have ratified the 2017 Nuclear Weapon Ban Treaty.

Based on the ideas of Tannenwald and Paul, this dissertation observes that a nuclear taboo emerged as a result of three factors. First, anti-nuclear civil society movements that expanded the discourse on nuclear weapons and labelled them as immoral. Second, normative power politics. Third, the categorisation of nuclear weapons as different from conventional weapons.

Civil society

In October 1949 citizens of Hiroshima staged the first rally to call for abolition of nuclear weapons. After the US hydrogen bomb test of 1954, a global grassroots movement spread across the globe, led by civil society groups such as the National Committee for a Sane Nuclear Policy and Campaign for Nuclear Disarmament in Britain. These groups organised demonstrations and meetings and engaged in civil disobedience including sailing into testing zones which generated significant media coverage (Tannenwald, 2005). The International Campaign to Abolish Nuclear Weapons (ICAN) played an instrumental role in advancing the 2017 Nuclear Weapon Ban Treaty. The anti-nuclear movement branded nuclear weapons as morally abhorrent and mobilized public opinion to put pressure of national leaders. They expanded the political discourse beyond military utility to include health, humanitarian and environmental effects.

The civil society movement against autonomous weapons is spearheaded by the Campaign to Stop Killer Robots. 12 of the 64 NGOs who comprise the Campaign have sent delegations to the CCW. Even if a ban falls though, the efforts to achieve this ban drawn attention to the issue. In July 2015, the Future of Life Institute (FLI) published an open letter calling for a ban on autonomous weapons which has been signed by some 3,900 AI and Robotics researchers including Elon Musk, Steve Wozniak, and DeepMind's Demis Hassabis (FLI, 2015). In November 2017 FLI released a short fictional film 'Slaughterbots' which depicts a terrorist attack enabled by autonomous microdrones; it has since been viewed over 2.5 million times. Hart suggests that certain classes of weapons are so morally repugnant that under no circumstances can their use ever be justified. From this perspective, debating the practicality of autonomous weapons or the feasibility of a ban largely misses the point. "Weapons are not banned because they can be", he argues, "but because there is something so morally abhorrent about them that nothing can justify their use" (Hart, 2017). Ban advocates should expand the discourse beyond civilian casualties in direct combat, proliferation to rogue actors, and moral concerns to emphasise risks of crisis instability and arms race escalation. However, this dissertation is wary that since the success of ICAN, powerful states have become more alert to the influence of civil society groups, and may pay more attention to restrict the effectiveness of these groups.

Power politics

After Stalin's death in 1953, Soviet leaders pursued a 'peace offensive' in which they used nuclear issues to exemplify their peaceful policy and characterise the West as aggressors (Tannenwald, 2005). Although this may have been propaganda (the Soviet Union was actively building its own nuclear weapons at the time), Soviet rhetoric tapped into global sentiment. The notion of 'rhetorical action' refers to the strategic use of norm-based arguments to advance self-interest. (Schimmelfenning, 2001). In a similar vein, Tannenwald suggests that the Soviets and other nonnuclear states engaged in a process of a process of 'strategic social construction' as they sought to shape of norms to suit their strategic interests (Tannenwald, 2005).

After the first CCW meeting in 2014, only 5 states supported a ban on LAWS; the number now stands at 26. With the exception of China, none of these states possess the capability to produce LAWS by themselves. In 2018, Brazil stated that "the idea that IHL can be embedded into LAWS is ethically and morally counterintuitive, and dangerous" (Brazilian Intervention GGE on LAWS, 2018, p. 2). Pakistan argued that "the absence of human intervention will make wars more inhumane. Regardless of the level of sophistication and programming, machines cannot replace humans in making the vital decision of taking another human's life" (Permanent Representative of Pakistan, 2018, p. 1). Therefore we can observe a growing number of states who oppose the development of LAWS. In line with this model of taboo creation, at the 2018 CCW meeting China stated its willingness "to negotiate and conclude a succinct protocol to ban the use of fully autonomous weapons systems" (China Mission Geneva, 2018); that same day the Chinese air force made known its aims to explore fully autonomous swarms of drones. Thus the juxtaposition of these announcements may be an attempt at strategic social construction. Kania observes that China has in the past sought to delegitimise adversaries with 'restriction through law', and prompts us to consider if China's objective is to place pressure on other states who must be more sensitive to public opinion (Kania, 2018).

Categorization

"The first step in stigmatizing an object or practice", argues Tannenwald, "is to redefine it as belonging in a separate category from otherwise similar objects" (Tannenwald, 2005, p. 17). The domestic politics of President Truman established nuclear weapons as separate from ordinary weapons. Rather than handing command to the military he put them under the authority of the Atomic Energy Commission, with only the US president having control of their use. Internationally, the establishment of the UN Atomic Energy Commission institutionalised the stigmatization of nuclear weapons. The UN further created the conceptual category of 'weapons of mass destruction' which encompasses atomic, radioactive, lethal chemical and biological weapons and sets them apart them from conventional weapons (Tannenwald, 2005). Paul observes that "the tradition seems have emerged due to the realisation that nuclear weapons are radically different from conventional weapons" (Paul, 2009, p. 14). One way in which nuclear weapons are different is the radiation hazards that they pose. Here we see similarities with other prohibited weapons. Price argues that "if there is one intrinsic quality of chemical weapons that seems to provide a plausible explanation for its prohibition, it is the association with poison" (Price, 2013, p. 80). In the case of nuclear and chemical weapons, "the mysterious attributes of the weapon might have played a role in the creation and persistence of the taboo" (Paul, 2009, p. 7). Altogether we see that weapons are stigmatised when they are set apart from conventional kinetic weapons.

As we have explored, a specific definition of LAWS required by legislation remains challenging to pin down. Nonetheless, when communicating with the media and general public, the humanitarian disarmament campaign works to portray autonomous weapons as separate from conventional weapons. The vocabulary employed by civil society – 'killer robots' – prompts a certain emotional response. The 'Slaughterbots' video arguably dramatizes the issue, which representatives from FLI believe is necessary because "serious discourse and academic arguments are not enough to get the message across" (Russel, Aguirre, Conn, & Tegmark, 2018). As our model predicts, we can observe that since 2018 civil society groups have begun to employ the term 'weapons of mass destruction' when engaging with the public. In an interview with CNN and in opinion pieces for The Guardian, Toby Walsh (who coordinated the boycott of KAIST) has argued that "autonomous weapons will be weapons of mass destruction" (Walsh, 2018). In January 2018, representatives from FLI stated that "our primary message has been consistent: Because they do not require individual human supervision, autonomous weapons are potentially scalable weapons of mass destruction" (Russel, Aguirre, Conn, & Tegmark, 2018). This varies from their July 2015 Open Letter, where the term 'weapons of mass destruction' is absent (FLI, 2015).

There is a fine line between stigmatising a target and stigmatising oneself. Carpenter observes that between 2007 and 2010 the nascent ban movement was inhibited by its association with science fiction; the issue was considered 'far out' and 'futuristic', thus those who proposed policy attention to it were stigmatised in the way that "anyone who talks about depleted uranium has got to be wearing a hat made out of tin foil" (Carpenter, 2016, p. 59). In the present day, Paul Scharre has criticised 'Slaughterbots' as "a piece of propaganda", arguing that the video undercuts a need for reasoned discussion (Scharre, 2017). This dissertation expresses concern that stigma will be attached to Terminator-style 'killer robots', permitting the creeping autonomisation of other systems. By becoming a focal point, this stigma may divert the normative gaze away from less visible but high-consequence risks (Tannenwald, 2008). Thus pro-ban advocates should categorise with caution and seek to mitigate against its permissive effects. Furthermore they should be ensure that the taboo stigmatises autonomous weapons specifically rather than the broader field of AI.

Altogether we can begin to detect the composite elements of an autonomous weapons taboo, especially with regard to the pathways of 'civil society' and 'power politics'. However it remains challenging to categorise autonomous weapons as separate from conventional weapons due to the widespread applications of AI and the difficulties of establishing deployment; they do not possess the 'all-or-nothing' character of nuclear weapons. Other considerations include the fact that nuclear weapons were difficult to develop, which slowed the pace of proliferation and 'bought time' for the taboo to develop. Tannenwald recognises that "had nuclear weapons spread quickly early on, the development of a taboo against their use would have been much less likely" (Tannenwald, 2008, p. 365). Finally we should be aware that the nuclear taboo was formed through "iterated behavior over time" (Tannenwald, 2008, p. 65). The notion of path dependence emphasises the importance of precedence and expectation; therefore the coming years remain crucial. This dissertation recommends the establishment of a new UN commission tasked with governing autonomous weapons. The CCW was a valuable starting point, a rigid forum focused on protecting civilians and combatants in direct conflict. However, as argued throughout this dissertation, effective governance must go beyond a ban. There is a need to broaden the discussion and the set of stakeholders who are engaged in it. This will facilitate the development of responsibility practices and enable continual dialogue, which is necessary given the rapid pace of technological development. The creation of a dedicated commission is also a normative enterprise which will embody shared values about the appropriateness of autonomous weapons.

Conclusion

For as long as humans have waged war, we have also sought to constrain it. Codes of conduct have been found in ancient Egypt, the Mahabharata, the book of Deuteronomy. In 1139 the Pope sought to ban the crossbow. Our current efforts to govern autonomous weapons carry on a lengthy tradition of figuring out how to kill each other, better.

Autonomous weapons present considerable difficulties for governance. AI promises to enhance numerous warfighting capabilities, flows seamlessly from the public into the military sphere, and blends unnoticed into conventional systems. Autonomous weapons generate exceptional risks, particularly if they exacerbate arms races, or destabilize networked military systems. Therefore this dissertation has proposed an approach to governance that begins with a legally binding ban on weapons that can search for targets, choose to engage, and deliver force without human input. This necessarily precise ban should be substantiated by forward- and backward-looking responsibility practices that cajole and command technical researchers to recognise that their algorithms have real-world consequences. Ongoing efforts to develop these two mechanisms should support and be supported by an autonomous weapons taboo.

All things considered, it is difficult to predict with a high level of confidence how successful the proposed mode of governance will be, or if and when unforeseen risks will arise. This dissertation has proposed an approach consisting of three mechanisms, but governance is an interactive process and we should continually refine and prioritize each mechanism. The proliferation of low-level autonomous weapons to rogue actors remains largely unsolved, primarily due to the availability of commercial drones; further research about how to address this risk is necessary. It is likely that the responses to this problem (and many others) will be AI-enabled and we should welcome but verify these solutions, paying attention to the impact of introducing ever-more autonomous systems to the landscape. More work must be done to predict the timeline of autonomous weapons development, and to investigate long-term risks areas including strategic stability and arms race dynamics. Many of the uncertainties in this dissertation – about the capabilities of AI, the rate of technological development, the magnitude of known and unknown risks – will not be resolved until events unfold. While the proposed mechanisms of governance may well prove to be ill-suited to the changing landscape, other recommendations are more timeless: learning from other fields, multi-stakeholder discussion, paying attention to less-visible but high-consequence outcomes.

The architecture of algorithms enables particular ways of working, frames decisions as they are made, and sways the points of human involvement. They are *codes* of conduct. But algorithms are not fixed, but built by and for humans in specific social, political and cultural contexts. It is up to us now, particularly those of us who shape laws and write code, to make decisions which will guide our relationship with conflict and each other. Let us make these decisions with compassion and reason.

Bibliography

Abbott, K. W., & Snidal, D. (2000). Hard and Soft Law in International Governance. *International Organization, 54*(3), 421-456.

Altmann, J., & Sauer, F. (2016). *Speed Kills: Why we Need to Hit the Brakes on "Killer Robots"*. Retrieved April 2018, from Duck of Minerva: http://duckofminerva.com/2016/04/speed-kills-why-we-need-to-hit-the-brakes-on-killer-robots.html

Amodei, D., Olah, C., Stienhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). *Concrete Problems in AI Safety.*

Arkin, R. (2007). *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture.* Georgia Institute of Technology, Mobile Robot Laboratory.

Arkin, R. (2009, July 24). *Teaching Robots the Rules of War*. Retrieved April 2018, from hplusmagazine: http://hplusmagazine.com/2009/07/24/teaching-robots-rules-war/

Arkin, R. (2018). In P. Scharre, *Army of none: autonomous weapons and the future of war.* New York: W. W. Norton & Company.

Armstrong, S., Bostrom, N., & Schulman, C. (2013). *Racing to the precipice: a model of artificial intelligence development.* Technical Report, Oxford University, Future of Humanity Institute.

Barnes, J. E., & Chin, J. (2018). *The New Arms Race in AI.* Retrieved from The Wall Street Journal: https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261

Bassiouni, M. C. (2002). *Accountability for Violations of International Humanitarian Law and Other Serious Violations.*

Bauman, Z. (2001). *Modernity and the Holocaust.* New York: Cornell University Press.

Bhuta, N., & Pantazopoulos, S.-E. (2016). Autonomy and uncertainty: increasingly autonomous weapons systems and the international legal regulation of risk. In *Autonomous Weapons Systems: Law Ethics, Policy.* Cambridge: Cambridge University Press.

Bhuta, N., & Pantazopoulos, S.-E. (2016). Autonomy and uncertainty; increasingly autonomous weapons systems and the international legal regulation of risk. In *Autonomous Weapons Systems: Law Ethics, Policy.* Cambridge: Cambridge University Press.

Bhuta, N., Beck, S., & Geib, R. (2016). Present future: concluding reflections and open questions on autonomous weapons systems. In *Autonomous Weapons Systems: Law, Ethics, Policy*. Cambridge: Cambridge University Press.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Boycott of Kaist. (2018, March). *Open Letter to Professor Sung-Chul Shin, President of KAIST from some leading AI researchers in 30 different countries.*

Retrieved April 2018, from Centre on Impact of AI and Robotics: https://www.cse.unsw.edu.au/~tw/ciair//kaist.html
Brazilian Intervention GGE on LAWS. (2018, April). *2018 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS).* Retrieved April 2018, from The United Nations Office at Geneva: https://www.unog.ch/80256EDD006B8954/(httpAssets)/44C33134FCF6BE6EC12582720056F0A2/$file/2018_LAWSGeneralExchange_Brazil.pdf
Campaign to Stop Killer Robots. (2013). *Campaign to Stop Killer Robots: The Problem*. Retrieved April 2018, from Campaign to Stop Killer Robots: https://www.stopkillerrobots.org/the-problem/
Campaign to Stop Killer Robots. (2017, December 22). *2017: A lost year for diplomacy*. Retrieved from Campaign to Stop Killer Robots: https://www.stopkillerrobots.org/2017/12/lostyear/
Campaign to Stop Killer Robots. (2018, April 11). Retrieved April 2018, from https://twitter.com/BanKillerRobots/status/984081170806857731
Campaign to Stop Killer Robots. (2018, April). *About Us*. Retrieved April 2018, from Campaign to Stop Killer Robots: https://www.stopkillerrobots.org/about-us/
Campaign to Stop Killer Robots. (2018). *Chronology*. Retrieved April 2018, from Campaign to Stop Killer Robots: https://www.stopkillerrobots.org/chronology/
Carpenter, C. (2014, May 14). *"Robot Soldiers Would Never Rape": Unpacking the Myth of the Humanitarian War-Bot*. Retrieved from Duck of Minerva: http://duckofminerva.com/2014/05/robot-soldiers-would-never-rape-unpacking-the-myth-of-the-humanitarian-war-bot.html
Carpenter, C. (2016). Rethinking the Political/ -Science- / Fiction Nexus: Global Policy Making and the Campaign to Stop Killer Robots. *Perspectives on Politics*.
Carvin, S. (2017). *Normal Autonomous Accidents: What happens when killer robots fail?* Working Paper.
Castells, M. (2011). A Network Theory of Power. *International Journal of Communication*, 773-787.
China Mission Geneva. (2018, April 13). CCW GGE on LAWS 1st Session. UN Digital Recordings Portal.
Committee on Research Standards and Practices to Prevent the Destructive Application of Biotechnology. (2004). *Biotechnology Research in an Age of Terrorism.* Washington DC: The National Academies Press.
Conference on Conventional Weapons. (2017). For consideration by the Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS). *Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects.* Geneva.

Conger, K., & Cameron, D. (2018, June 3). *Google is Helping the Pentagon Build AI for Drones*. Retrieved from Gizmodo: https://gizmodo.com/google-is-helping-the-pentagon-build-ai-for-drones-1823464533

Corn, G. S. (2016). Autonomous weapons systems: managing the inevitability of 'taking the man our of the loop'. In *Autonomous Weapons Systems: Law, Ethics. Policy* (pp. 209-242). Cambridge: Cambridge University Press.

Crootof, R. (2016). A Meaningful Floor for "Meaningful Human Control". *Temple International & Comparative Law Journal, 30*.

Cummings, M. (2006). Automation and Accountability in Decision Support System Interface Design. *Journal of Technology Studies*, 23-31.

Cummings, M. L. (2017). *Artificial Intelligence and the Future of Warfare.* Chatham House.

DARPA. (2018, March 31). *DARPA News and Events.* Retrieved from Defense Advanced Research Projects Agency: https://www.darpa.mil/news-events/2018-03-30

Defense Innovation Board. (2018). *Recommendation* . Retrieved from Defense Innovation Board: http://innovation.defense.gov/Recommendations/

Defense Science Board. (2016). *Summer Study on Autonomy.* Washington: Department of Defense.

Deputy Secretary of Defense. (2018, April 26). *Establishment of an Algorithmic Warfare Cross-Functional Team (Project Maven).* Retrieved from Government Executive: https://www.govexec.com/media/gbc/docs/pdfs_edit/establishment_of_the_awcft_project_maven.pdf

Docherty, B. (2012). *Losing Humanity: The Case against Killer Robots*. Retrieved April 2018, from Human Rights Watch: https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots

Docherty, B. (2015). *Mind the Gap: The Lack of Accountability for Killer Robots*. Retrieved April 2018, from Human Rights Watch: https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots

Dunlap, C. (2015, April 3). *Guest Post: To Ban New Weapons or Regulate Their Use?* Retrieved April 2018, from Just Security: https://www.justsecurity.org/21766/guest-post-ban-weapons-regulate-use/

Dunn, J. (2017, May 23). *Drones are growing rapidly, regardless of what the government does*. Retrieved April 2018, from Business Insider: http://uk.businessinsider.com/drone-sales-in-us-chart-2017-5

Dvorsky, G. (2015, July 29). *Why the Argument Against a Ban on Autonomous Killer Robots Falls Flat*. Retrieved from io9: https://io9.gizmodo.com/why-the-argument-against-a-ban-on-autonomous-killer-rob-1720852797

Evans, H. (2018, April 13). *Too Early for a Ban: The U.S. and U.K. Positions on Lethal Autonomous Weapons Systems*. Retrieved April 2018, from Lawfare: https://lawfareblog.com/too-early-ban-us-and-uk-positions-lethal-autonomous-weapons-systems

FAT ML. (2016). *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*. Retrieved April 2018, from Fairness, Accountability, and Transparency in Machine Learning: https://www.fatml.org/resources/principles-for-accountable-algorithms

Fearon, J. (1995). Rationalist Explanations for War. *International Organization*, 379-414.

FLI. (2015). *Autonomous Weapons: An open letter from AI & Robotics reserchers*. Retrieved April 2018, from Future of Life Institute: https://futureoflife.org/open-letter-autonomous-weapons/

Good, I. J. (1964). Speculations Concerning the First Ultraintelligent Machine.

Goose, S., & Wareham, M. (2017, January 5). *The Growing International Movement Against Killer Robots*. Retrieved 2018 April, from Human Rights Watch: https://www.hrw.org/news/2017/01/05/growing-international-movement-against-killer-robots

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). *When Will AI Exceed Human Performance? Evidence from AI Experts.*

Gramone, J. (2016, April 27). *U.S. Troops Should Not Be Sent Into Fair Fights, Dunford Says*. Retrieved April 2018, from U.S. Department of Defense: https://www.defense.gov/News/Article/Article/744390/us-troops-should-not-be-sent-into-fair-fights-dunford-says/

Hart, R. (2017, April 24). *The case for banning autonomous weapons rests on morality, not practicality*. Retrieved April 2018, from Bulletin of the Atomic Scientists: https://thebulletin.org/case-banning-autonomous-weapons-rests-morality-not-practicality10707

ICJ. (1996). *Legality of the threat or use of nuclear weapons.* International Court of Justice.

ICRAC. (2013, October 16). *Computing experts from 37 countries call for ban on killer robots*. Retrieved March 2018, from International Committee for Robot Arms Control: https://www.icrac.net/call/

IEEE Global Initiative . (2016). *Ethically Aligned Design Version One - For Public Design.* IEE.

International Criminal Court. (1998). *Rome Statute, art. 25.*

Jain, N. (2016). Autonomous weapons systems: new frameworks for individual responsibility. In *Autonomous Weapons Systems: Law, Ethics, Policy.* Cambridge: Cambridge University Press.

Johnson, D. G. (2015). Technology with No Human Responsibility? *Journal of Business Ethics*, 707-715.

Johnson, D., & Noorman, M. (2014). Responsibility Practices in Robotic Warfare. *Military Review*, 12-22.

Judgement of the Nuremberg International Military Tribunal. (1947). *American Journal of International Law, 41*.

Kalmanovitz, P. (2016). Judgement, liability and the risks of riskless warfare. In N. Bhuta, S. Beck, R. Geib, H.-Y. Liu, & C. Kreb (Eds.), *Autonomous Weapons Systems* (pp. 145-163). Cambridge: Cambridge University Press.

Kania, E. (2018, April 17). *China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems*. Retrieved from Lawfare : https://www.lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems

Kaushal, M., & Nolan, S. (2015). *Understanding Artificial Intelligence.* The Brookings Institution.

Lee, P. (2018, April 11). *Drone will soon decide who to kill*. Retrieved from The Conversation: https://theconversation.com/drones-will-soon-decide-who-to-kill-94548

Lessing, L. (2000). *Code Is Law.* Retrieved April 2018, from Havard Magazine: https://harvardmagazine.com/2000/01/code-is-law-html

Letter to Google C.E.O. (2018). *Microsoft Word - Letter to Google C.E.O*. Retrieved April 2018, from The New York Times: https://static01.nyt.com/files/2018/technology/googleletter.pdf

Lewis, J. (2015). The Case for Regulating Fully Autonomous Weapons. *The Yale Law Journal, 124*(4).

Liu, H.-Y. (2016). Refining responsibility: differentiating two types of responsibility issues raised by autonomous weapons systems. In *Autonomous Weapons Systems: Law, Ethics, Policy* (pp. 325-344). Cambridge: Cambridge University Press.

Marchant, G. E. (2011). Addressing the Pacing Problem. In A. M. Cutter, B. Gordijn, G. E. Marchant, & A. Pompidou (Eds.), *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight.* Springer.

Markoff, J. (2016, May 11). *Pentagon Turns to Silicon Valley for Edge in Artificial Intelligence*. Retrieved from The New York Times: https://www.nytimes.com/2016/05/12/technology/artificial-intelligence-as-the-pentagons-latest-weapon.html

McHarg, I. (1969). Design With Nature. *The Natural History Press*, 67-70.

Noorman, M. (2013). Responsibility Practices and Unmanned Military Technologies. *Sci Eng Ethics*, 809-826.

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

Office of Science Policy. (2012). *United States Government Policy for Oversight of Life Sciences Dual Use Research of Concern*. Retrieved April 2018, from Dual

Use Research of Concern: http://www.phe.gov/s3/dualuse/Documents/us-policy-durc-032812.pdf
Paul, T. (2009). *The Tradition of Non-Use of Nuclear Weapons.* Stanford: Stanford University Press.
Permanent Representative of Pakistan. (2018, April 9). *Statement by Ambassador Farukh Amil Permanent Representative of Pakistan.* Retrieved April 2018, from Reaching Critical Will: http://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/statements/9April_Pakistan.pdf
Perrigo, B. (2018, April 9). *A Global Arms Race for Killer Robots Is Transforming the Battlefield*. Retrieved 2018 April, from TIME: http://time.com/5230567/killer-robots/
Pierson, P. (2000). Increasing Returns, Path Dependence, and the Study of Politics. *American Political Science Review*, 251-267.
Pinker, S. (2018). *Enlightenment Now.* New York: Penguin.
Price, R. (2013, January 22). *How Chemical Weapons Became Taboo*. Retrieved from Foreign Affairs: https://www.foreignaffairs.com/articles/syria/2013-01-22/how-chemical-weapons-became-taboo
Prosecuter v. Jean-Pierre Bemba Gombo, ICC-01/05-01.08 (Pre-Trial Chamber II June 15, 2009).
Prosecuter v. Strugar. (2008). Case No. IT-01-42-A, Judgement (Appeals Chamber). ICTY.
Roff, H. (2016, Febuary 9). *Autonomous weapons, civilian safety, and regulation versus prohibition*. Retrieved April 2018, from Bulletin of the Atomic Scientists: https://thebulletin.org/autonomous-weapons-civilian-safety-and-regulation-versus-prohibition
Roff, H. (2018). *Autonomy, Robotics & Collective Systems*. Retrieved April 2018, from Arizona State University Global Security Initiative: https://globalsecurity.asu.edu/robotics-autonomy
Roff, H. (2018). *Dataset: Survey of Autonomous Weapons Systems.* Retrieved from Arizona State University Global Security Initiative: https://globalsecurity.asu.edu/robotics-autonomy
Russel, S., Aguirre, A., Conn, A., & Tegmark, M. (2018, January 23). *Why You Should Fear 'Slaughterbots' - A Response*. Retrieved April 2018, from IEEE Spectrum: https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/why-you-should-fear-slaughterbots-a-response
Saalman, L. (2018, April 24). *Fear of false negatives: AI and China's nuclear posture*. Retrieved from Bulletin of the Atomic Scientists: https://thebulletin.org/military-applications-artificial-intelligence/fear-false-negatives-ai-and-china%E2%80%99s-nuclear-posture?platform=hootsuite

Salmon, F., & Stokes, J. (2010, Obtober 27). *Algorithims Take Control of Wall Street*. Retrieved April 2018, from Wired: https://www.wired.com/2010/12/ff-ai-flashtrading/

Saxon, D. (2016). A human touch: autonomous weapons, DoD Directive 3000.09 and the intepretation of "appropriate levels of human judgement over the use of force. In *Autonomous Weapons Systems: Law, Ethics, Policy.* Cambridge: Cambridge University Press.

Saxon, D. (2016). A human touch: autonomous weapons, DoD Directive 3000.09 and the intepretation of "appropriate levels of human judgement over the use of force. In *Autonomus Weapons Systems: Law, Ethics, Policy.* Cambridge : Cambridge University Press.

Scharre, P. (2017, Dec 22). *Why You Shouldn't Fear 'Slaughterbots'*. Retrieved April 2018, from IEEE Spectrum: https://spectrum.ieee.org/automaton/robotics/military-robots/why-you-shouldnt-fear-slaughterbots

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

Scherer, M. (2016). Regulating AI systems: Risks, Challenges, Competencies and Strategies. *Harvard Journal of Law & Technology, 29*(2).

Schimmelfenning, F. (2001). The Community Trap: Liberal Norms, Rhetorical Action, and the Eastern European Enlargement of the European Union. *International Organization*.

Schmidt, E. (2017, Nov). Eric Schmidt Keynote Address - Artifical Intelligence and Global Security. Center for a New American Security (CNAS).

Select Committee on Artificial Intelligence. (2018). *AI in the UK: ready, willing and able?* London: House of Lords.

Sending, O. J., & Neumann, I. B. (2006). Governance to Governmentality: Analyzing NGOs, States, and Power. *International Studies Quarterly*, 651-672.

Shelton, D. (2005). *Remedies in International Human Rights Law.* Oxford: Oxford University Press.

Simmons, B. (2010). Treaty Compliance and Violation. *Annual Review of Political Science*.

Singer, P., & Friedman, A. (2014). *Cybersecurity and Cyberwar.* Oxford: Oxford University Press .

Statement by the Czech Republic. (2015, April 13). *2015 Meeting of Experts on Lethal Autonomous Weapons Systems Convention on Certain Conventional Weapons General Debate.* Retrieved April 2018, from Permanenet Mission of the Czech Republic to the United Nations Office and other International Organizations: https://www.mzv.cz/public/29/e/7d/1448252_1299062_CZ_statement_general_debate_LAWS_ver2_1.pdf

Syrian Observatory for Human Rights. (2018, December 31). *About 60 thousand were killed in 2016 and we still wait for an international trial for the criminals*. Retrieved April 2018, from Syrian Observatory for Human Rights: http://www.syriahr.com/en/?p=58114

Tannenwald, N. (2005). Stigmatizing the Bomb: Origins of the Nuclear Taboo. *International Security*, 5-49.

Tannenwald, N. (2008). *The Nuclear Taboo.* Cambridge: Cambridge University Press.

Tucker, P. (2018, April 19). *The Pentagon Is Building an AI Product Factory*. Retrieved from Defense One: http://www.defenseone.com/technology/2018/04/pentagon-building-ai-product-factory/147594/?oref=d-river

UN General Assembly. (2005, December 16). *Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law*. Retrieved April 2018, from United Nations Human Rights: http://www.ohchr.org/EN/ProfessionalInterest/Pages/RemedyAndReparation.aspx

United Nations Office for Disarmament Affairs. (2014). *Convention on Certain Conventional Weapons.* Retrieved from United Nations Office for Disarmament Affairs: https://unoda-web.s3-accelerate.amazonaws.com/wp-content/uploads/assets/publications/more/ccw/ccw-booklet.pdf

UNOG. (2018). *Latest Information*. Retrieved April 2018, from The United Nations Office at Geneva: https://www.unog.ch/80256EE600585943/(httpPages)/87CF9BFD24A8D05FC1257574004B285B?OpenDocument

Vilmer, J.-B. J. (2016, September 27). *Autonomous Weapons Diplomacy: The Geneva Debates*. Retrieved April 2018, from Ethics & International Affairs: https://www.ethicsandinternationalaffairs.org/2016/autonomous-weapon-diplomacy-geneva-debates/

Walsh, T. (2018, April 6). *It's not too late to save the world from killer robots*. Retrieved April 2018, from The Guardian: https://www.theguardian.com/commentisfree/2018/apr/06/its-not-too-late-to-save-the-world-from-killer-robots

Weinberg, A. (1975). Maturity and the future of nuclear energy. *American Scientist.*

WHO. (2018). *Dual Use Research of Concern*. Retrieved from World Health Organization: http://www.who.int/csr/durc/en/