The Hague Centre for Strategic Studies

HCSS Security

Towards Responsible Autonomy

The Ethics of Robotic and Autonomous Systems in a Military Context

September 2019

This paper is part of the HCSS project on "The Military Application of Robotic and Autonomous Systems: What, Why, How and Under What Conditions", conducted for and financed by the Netherlands Ministry of Defence. This paper is published separately from the forthcoming capstone document on the project. Responsibility for the content and for the opinions expressed rests solely with the authors; publication does not constitute an endorsement by the Netherlands Ministry of Defence.

Authors: Esther Chavannes and Amit Arkhipov-Goyal

Project Team: Patrick Bolder, Bianca Torossian, Michel Rademaker and Frank Bekkers

Quality Assurance: Dr. Arthur van der Wees (Arthur's Legal), Dr. Elke Schwarz (Queen Mary University London), Dr. Marijn Hoijtink (Vrije Universiteit Amsterdam), Gen. BD. Tom Middendorp (former Chief of Defence of the Netherlands Armed Forces) and Lt. Col. Ir. S.A.N. Mevissen (Royal Netherlands Army)

The authors express their gratitude to Lt. Col. Ir. E.P. Verdiesen (Royal Netherlands Army/TU Delft) and Ms. Joanna van der Merwe (Leiden University) for their valuable feedback in the review process.

Executive Summary

The current ethical debate on robotic and autonomous systems (RAS) is often dominated by relatively extreme narratives surrounding a total ban on 'killer robots' (a euphemism for lethal autonomous weapon systems) entirely. While there are many valid ethical concerns, current discussions on RAS have sidelined nuances that have critical implications for deciding how to introduce RAS in a military context. The brewing AI arms race and the diffusion of cheap, technologically advanced systems among state and non-state actors compels countries to adopt RAS. This is due to not only the prospect of lagging behind allies, but more so responding to adversaries using RAS to create a significant military advantage and escalation dominance.

With the perspective that RAS will be further incorporated in the military context, this paper presents a balanced discussion of the key overarching topics of ethical concern that arise from the introduction of RAS: human agency, human dignity, and responsibility. In short, these topics of debate 1) concern the ability of humans to retain control over systems; 2) weigh the positive and negative ways in which RAS in a military context contribute to respect for human dignity; and 3) assess the shortcomings of present responsibility structures for deploying RAS.

(Semi-)autonomous systems have been in operation for over four decades. As systems become increasingly independent, with the capability to perform their own calculations rather than just being bound by a set of rules, the concern for uncontrollable or unexplainable robots has arisen. In reality, however, the advent of systems matching human intelligence is unlikely to be achieved in the coming decade. This redirects the focus to more functional challenges that have an impact on ethical behavior, such as the design of systems, decreasing understanding of algorithmic calculations, and cognitive challenges arising from human–machine teaming. To address this, the paper presents a three-part framework through which to identify human control within a system: through the life cycle of RAS, through RAS' sub-system functions, and through the observe-orient-decide-act (OODA) loop.

On the topic of human dignity, International Humanitarian Law is crucial in assessing ethical use of RAS. Coupled with discussions on how to legally and ethically integrate RAS into the military, there is a fierce debate between arguments that increased use of RAS may either aid or undermine respect for human dignity. Aside from International Humanitarian Law, which can govern users', operators' and commanders' behavior with RAS in warfare, elements of both civil and criminal law may be relevant in addressing questions of responsibility and accountability for the actions of RAS. Assessing accountability is important, both for cases of active wrongdoing, and for identifying and addressing mistakes that may occur in the further integration of RAS into the military. Care should be taken to anticipate future risks of increased

autonomy and to address the possibility of a resulting accountability gap ahead of time. New frameworks or a different use of existing bodies of law may be necessary, for example by considering institutionalizing a 'system of control' involving all relevant actors throughout the entire life and use cycles of RAS.

The collective decision to deploy RAS is that of the public, the government and the armed forces, but one that must be made on an informed basis. As a result of this study, several recommendations are offered to the Netherlands Ministry of Defence and the Royal Netherlands Army (for the complete list of recommendations, see Chapter 6):

- The fundamental principle is to work with 'ethics by design', whereby ethical considerations are incorporated in the use case identification, system design, validation, manufacturing, and testing processes, rather than solely in the operation stage of the system life cycle;
- Build understanding of the system performance and behavior through the involvement of end users as early as in the design and testing stages, with the end goal to be for operators, supervisors and commanders to be able to trace, understand and predict the system's decision-making process;
- Develop best practice guidelines for (1) the outsourcing of the development process to external contractors; and (2) interoperability frameworks with technologically advanced allied armed forces that co-deploy RAS;
- Identify within what sub-system functions of RAS increasing automation and autonomy will present benefits to the military without eliciting major ethical concerns, e.g., movement controls, sensory controls and computer vision;
- Program core rules of engagement (ROEs) with International Humanitarian Law principles embedded in system design, along with an open architecture to introduce mission-specific ROEs by mission command;
- Improve transparency on the uses and contexts of use of RAS in the military domain with the general public.
- Create an institutional culture of shared accountability concerning all actors involved throughout the RAS life cycle.

While the recommendations presented above are not the sole solutions to existing ethical challenges, they do present pathways forward in the incorporation of ethical principles in RAS within the Royal Netherlands Army. Rapid advances in computational power and data generation are paving the way for exponential growth in the sophistication of RAS, making this is a salient issue for both the Netherlands Ministry of Defence and the Royal Netherlands Army, as well as governments and military forces elsewhere. The above recommendations are therefore presented with a distinct sense of urgency.

Contents

1 | Introduction

Throughout history, the invention of new military technologies has fundamentally changed how wars are fought.[1] The introduction of Robotic and Autonomous Systems (RAS) is no different and has led to renewed concerns over ethical issues associated with the use of new technologies in military forces. This is particularly salient in the context of autonomous weapon systems (AWS).[2] Supervised and fully autonomous systems have been in operation for several decades in over thirty countries. These have previously raised little ethical concern, even with their high degree of autonomy and often lethal designation, such as the Israeli Harpy and the US Tomahawk Anti-Ship Missile,[3] the latter of which was already withdrawn from service in the US Navy in the 1990s.[4] Their application is most frequent in cases where engagements supersede human decision-making and reaction times. The proliferation of increasingly autonomous systems has been expanding exponentially, with at least sixteen countries and several non-state actors, such as Hezbollah in Lebanon and Houthi rebels in Yemen, being in possession of armed unmanned aerial vehicles (UAV).[5]

Definitions

Throughout this paper, a differentiation is made between Robotic and Autonomous Systems (RAS) and autonomous weapons systems (AWS), according to the following basic definitions of RAS and AWS:

Robotic and Autonomous Systems (RAS)

RAS is an accepted term in academia and the science and technology community and highlights the physical (robotic) and cognitive (autonomous) aspects of these systems. For the purposes of this concept, 'RAS' is a framework to describe systems with a robotic element, an autonomous element, or, more commonly, both.[6]

[1] Banta, "'The Sort of War They Deserve'?"

[2] Some scholars indicate that this balancing difficulty is less attributable to technology alone, but as much so to governments' choices, such as the development toward presuming some right to anticipatory self-defense, like in the US' drone campaigns. It could even be argued that this ethical discussion is not new per se, and mirrors the one concerning the development of air power in World War II, see Boyle, "The Legal and Ethical Implications of Drone Warfare."

[3] The Tomahawk Anti-Ship Missile (TASM) should not be confused with the Tomahawk Land Attack Missile (TLAM), which is still in service to this date and operates under a different set of parameters. Scharre, *Army of None: Autonomous Weapons and the Future of War*, 47–49.

[4] Scharre, 47–49.

[5] Scharre, 102–3.

[6] The definition is borrowed in full from Feickert et al., "U.S. Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress."

Autonomous Weapon Systems (AWS)

These are weapon systems that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that can select and engage targets without further human input after activation but are designed to allow human operators to override the operation of the weapon system.[7] Variation upon this US Department of Defense definition exists internationally, and characteristics such as 'intelligence', the possibility to learn or adapt, or a level of unpredictability are sometimes included.[8] Some, for example, consider the ability to search for targets by maneuvering intelligently through an environment to be a feature of an autonomous weapon system.[9]

For the purposes of this paper, RAS is an all-encompassing term that refers to systems with any degree of autonomy and any military designation, whether for communication, logistics, reconnaissance, weapons delivery or otherwise, meaning the definition is inclusive of, but is not limited to AWS.

Ethical Controversy in the Use of Robotic and Autonomous Systems

Arguments for the importance of RAS are numerous, and primarily encompass the ensuing technological arms race, diffusion of military power, and societal expectations of lower numbers of civilian casualties.[10] Concern for the adversarial development of RAS is one of the more frequently cited reasons for a state's own development of such systems.[11] While normative actors such as the Netherlands do not seek to delegate absolute authority to machines,[12] adversarial state and non-state actors in possession of autonomous systems may gain a competitive advantage and, as a result, present a security risk. Among state actors, this is developing into an 'AI arms race', where countries feel the need to develop AI-driven systems because other states are or may be doing so. The accessible nature of AI and robotic hardware makes such systems an option for both individuals and groups, creating a whole host of security risks.[13]

[7] The definition is borrowed in full from Feickert et al.

[8] Kania, "China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems," April 17, 2018; Ekelhof, "Autonome Wapensystemen: Wat We Moeten Weten over de Toepassing van Het Humanitair Oorlogsrecht En de Menselijke Rol in Militaire Besluitvorming," 194; Chairperson of the Informal Meeting of Experts, "Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)."

[9] Scharre, *Army of None: Autonomous Weapons and the Future of War*, 123.

[10] Scharre, 95, 117, 134.

[11] Jones et al., "Managing the New Threat Landscape: Adapting the Tools of International Peace and Security," 18.

[12] Netherlands Advisory Council on International Affairs, "Autonomous Weapon Systems: The Need for Meaningful Human Control."

[13] Jones et al., "Managing the New Threat Landscape: Adapting the Tools of International Peace and Security."

Beyond technological and strategic competition, Western societies have also come to expect fewer civilian and soldier casualties in warfare due to technological advancement of the arsenals of nation-states. While individual civilian casualties resulting from drone strikes are questioned today, just 75 years ago during the Second World War, nation-states were carpet-bombing cities, with hundreds or thousands of deaths resulting from individual air campaigns.[14] As the societal tolerance for civilian casualties decreases, the need for advanced systems of defense and precision-guided weapons becomes more apparent.

The International Committee of the Red Cross (ICRC) has concluded from various opinion polls in over fifty countries that, when it comes to autonomous weapon systems, the most prominent ethical concerns are those regarding "loss of human agency in decisions to use force—decisions to kill, injure and destroy—, loss of human dignity in the process of using force, and erosion of moral responsibility for these decisions."[15] When put into a broader perspective, to also include non-weapon systems, the same ethical concerns remain relevant. These concerns touch upon questions of human agency, human dignity, and responsibility. Although some overlap can be found between these three overarching ethical challenges, together they cover the most relevant concerns associated with the adoption of RAS by armed forces. Considering the recent developments in RAS proliferation, these systems' utility for militaries worldwide means that RAS are already in use, and capabilities will continue to be developed further.

The objective of this paper is to present existing ethical challenges in the use of RAS, highlight the issues that have previously received little attention, and discuss pathways for the ethical integration of RAS in the Royal Netherlands Army (RNLA).

This paper will not only consider lawfulness and lawful use of RAS under International (Humanitarian) Law, but will also delve into the moral discussion on the use of RAS and human dignity, as well as questions of understanding, bias, accountability and responsibility. Within this wide discussion on the ethical, legal and social ramifications of these technological developments,[16] the structure of this paper is based on the following questions:

[14] Scharre, *Army of None: Autonomous Weapons and the Future of War*, 282.

[15] While the ICRC is just one of the stakeholders in the discussion on the ethics of RAS, and the results of opinion surveys have limitations, the ICRC has guided the humanitarian perspective of warfare throughout modern history. Following the Geneva Conventions, the ICRC remains the primary normative actor focused on maintaining humanity in warfare and, as a result, presents the fundamental humanitarian concerns arising from the use of RAS in a military context; see Davison, "Autonomous Weapon Systems"; International Committee of the Red Cross ICRC, "Ethics and Autonomous Weapon Systems," 21.

[16] Floridi, "What the Near Future of Artificial Intelligence Could Be."

- How and to what extent can and should human agency, and human control in particular, be retained in the operation of RAS?
- What does it take to maintain human dignity in the operation of RAS?
- Who is, or should be, responsible for the actions and outcomes of the use of RAS?

These guiding questions and the three aforementioned key ethical issues form the structure of this paper. After presenting first in Chapter 2 the background and recent developments relevant to discussions on RAS, Chapter 3 covers the topic of human agency, and human control in particular, in the use of RAS. It proposes and discusses a framework for how to assess what constitutes 'meaningful human control', and it lays out challenges such as explainability, self-learning abilities, and the complexities of human–machine teaming. Chapter 4 on human dignity positions the ethical debate on RAS, and on AWS in particular, in the context of existing frameworks of International Humanitarian Law. It also discusses the current debate on how AWS could undermine or enhance human dignity when used in a military context. Chapter 5 on accountability and responsibility presents the complexity of holding states and/or (groups of) individuals responsible in the search for legal accountability. It then addresses the phenomenon of an 'accountability gap' formed by the outpacing of laws and social norms by technological progress, as well as possible ways to mitigate this challenge. Finally, Chapter 6 concludes the discussion and provides recommendations for the RNLA at the strategic and operational levels.

Research Process

Current debates surrounding RAS have yet to result in concrete guidance for tackling the ethical challenges of RAS integration that the armed forces are faced with. This paper seeks to address this in order to guide further discussions on the introduction of these systems within the Royal Netherlands Army, as well as in partner states. This paper draws its conclusions from an extensive literature review and an expert session hosted by *The Hague* Centre for Strategic Studies (HCSS) on June 13th, 2019. The expert session was based on four fictional scenarios (attached in Appendix A) that were created to test the cognitive boundaries on the ethical challenges posed by the use of RAS in particular contexts. The session involved military, legal, technical, and ethics experts from the Netherlands Ministry of Defence, Royal Netherlands Army, Netherlands Organisation for Applied Scientific Research (TNO), academic institutions, non-governmental organizations and the private sector. The participants were assigned into four groups, with each group completing all four scenarios. This way, four sets of perspectives were recorded for each scenario and were subsequently noted in the session summaries provided in Appendix B.

2 | Background and Recent Developments

It is critical to note that most applications of RAS in a military setting are not for lethal designations and span a variety of roles, including surveillance, logistics, medical support, maintenance, communication and engineering.[17] It is estimated that out of the known 500 RAS in operation today worldwide, 30% are designated for the use of force, within which 55% are used for defensive and 45% for offensive purposes. This means that 14% of all systems currently deployed have a lethal offensive component.[18] While current debates surrounding the ethics of RAS tend to focus on this small portion of systems, this paper discusses relevant ethical considerations for many different possible RAS applications. Lethal systems will only be on the foreground in Chapter 4 when International Humanitarian Law is discussed.

Non-lethal systems and applications continue to demonstrate landmark achievements, such as the US Navy's Autonomous Aerial Cargo/Utility System, which autonomously determined an improvised landing zone and carried out an autonomous landing in 2014 (see Figure 1).[19] A year later, the US X-47B unmanned aerial vehicle (UAV) conducted the first fully autonomous air-to-air refueling.[20] In the summer of 2019, the Netherlands Army 13th Brigade trained with two THeMIS combat support unmanned ground vehicles (UGV) in Scotland, introduced as logistic support for deployed troops (see Figure 2).[21]

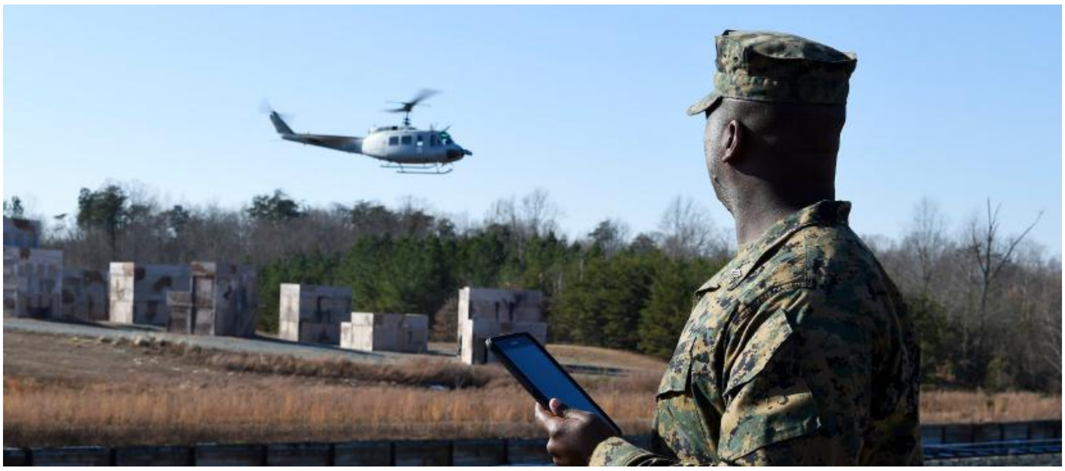

Figure 1. A UH-1 Huey equipped with the Autonomous Aerial Cargo/Utility System (AACUS). Photo: John F. Williams/US Navy

[17] Torossian et al., "Paper on the Military Applicability of Robotic and Autonomous Systems," 15.

[18] Percentage calculations are based on: Torossian et al., 15–16. While some systems have distinct applications, others have multiple purposes, such as unmanned aerial vehicles capable of acting both as surveillance and weapons delivery systems, meaning this can affect the abovementioned figures.

[19] Scharre, *Army of None: Autonomous Weapons and the Future of War*, 17.

[20] JASON, The MITRE Corporation, "Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD," 4.

[21] "Milrem Robotics Delivered Two THeMIS UGVs to the Dutch Army."

Figure 2. RNLA training with THeMIS UGV in Scotland, 2019. Photo: Milrem Robotics

Similarly to RAS, Artificial Intelligence (AI) algorithms that power autonomous systems have been in development since the 1950s, and their increasing abilities are driven by a newfound capacity to collect, store and process mass amounts of various types of data.[22] It is therefore critical to stress that neither RAS, nor the AI powering them, are new. Rather, due to the advancements of computational power and the increased amounts of data ('big data') generated in the last decade, capabilities are now expanding far beyond initial abilities, increasingly beyond the levels of human cognition. Breakthroughs have been made in the use of various deep learning (DL) algorithms such as deep neural networks (DNN).[23] An example of this is the error rate of visual object recognition, which through the use of such a neural network decreased from 25% to 3% between 2011 and 2015, compared to a fixed human error rate of 5%.[24] While not indicative of use in complex or uncontrolled environments, this demonstrates that AI can already supersede human abilities in certain contexts and will continue to do so in the future. The upward trend in such systems' use and their ability to carry out more functions independently, often better than their human counterparts, is causing shifts in perception of and attitudes towards RAS.[25]

[22] Scott et al., "Modeling Artificial Intelligence and Exploring Its Impact"; Spiegeleire, Maas, and Sweijs, *Artificial Intelligence and the Future of Defense*, 31–39.

[23] Essentially, the learning process of a deep neural network is a process through which a large data set (e.g., of images to be recognized and classified) combined with high computing power makes it possible to filter through all recognizable elements in the data set in order to build a new model. Whereas older machine learning techniques required you to first build a model to recognize the data points for what they were, with deep leaning the computer model 'teaches' itself what the defining features of all the different objects/images/etc. in the data set are. A simple example of this would be a data set of dog photos, with each photo labeled as the breed of the dog in question, from which a DNN trains to establish the features that make up a dog's breed in a way that will allow the DNN to recognize the breeds of dogs in new photos.

[24] JASON, The MITRE Corporation, "Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD," 9.

[25] McLean, "Drones Are Cheap, Soldiers Are Not"; "Dilbert at War."

3 | Human Agency

This chapter presents the technical complexity of RAS, and it discusses diverging approaches to establishing the degree of human control over RAS as well as the impact this has on human–machine teaming. The chapter first presents definitions of key concepts, explains the varying degrees of autonomy that are possible, and dissects the functional complexity of RAS across the life cycle, sub-system functionality, and the observe-orient-decide-act (OODA) loop. This is followed by a discussion on the determination of the acceptable degree of human control and the main factors affecting this determination, namely the explainability and predictability of machines, self-learning abilities, and software updates. Building on these factors, the chapter discusses challenges in human–machine teaming that were identified in the literature review and the expert session. These are automation bias and complacency, distinction between trust or knowledge of systems and providers, interoperability issues, and the anthropomorphizing (i.e., humanizing) of the machines. While these topics are not the sole factors that elicit ethical considerations in the use of RAS, they were selected as a result of prominence in academic literature and among experts at the HCSS scenario-based expert session. The latter enabled the discussion of issues which may be underrepresented in academic literature and only arise in certain contexts.

Maintaining human agency, particularly in the context of AWS, is one of the most contentious issues of debate with respect to the integration of RAS in the military domain. Human agency is a concept that encompasses "self-control, morality, memory, emotion recognition, planning, communication and thought."[26] It includes "features of self-awareness, self-consciousness and self-authorship," and as a result relates to moral agency and affects the attribution of responsibility.[27] Human control, also referred to throughout this paper as 'meaningful human control' (MHC), is an operational component of human agency, which distinguishes between human and artificial decision-making processes.[28] The term has been adopted by a number of state and non-state actors to frame the discussion on human control over and autonomy in weapon systems.[29] Although the discussion on MHC primarily concerns

[26] Gray, Gray, and Wegner, "Dimensions of Mind Perception."

[27] European Group on Ethics in Science and New Technologies (EGE), *Statement on Artificial Intelligence, Robotics and "autonomous" Systems*.

[28] MHC interrelates with "effective control", a prerequisite in public international law for legal liability and unlawful conduct. In the context of the use of RAS/AWS, the term is used alongside "effective command" to determine state responsibility. This is discussed further in Chapter 4. European Group on Ethics in Science and New Technologies (EGE); "Killer Robots and the Concept of Meaningful Human Control."

[29] MHC is not used universally, and controversy over the definition primarily centers around the degree to which there is a 'human in the loop', meaning, the degree to which a human is involved in the operating and/or decision-making process of the system. See United Nations Institute for Disarmament Research (UNIDIR, "The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward."

the operation of RAS, suggestions have been made that for control at the operational stage to be meaningful, there must also be sufficient control throughout the wider life cycle of a system, incorporating it into the design, procurement, testing, and decommissioning stages.[30]

3.1 Human Control

Establishing an agreed upon definition of meaningful human control is hindered by the arbitrary nature of the acceptable degree of control and by the varying approaches to the concept of MHC among the nations that lead the development and application of RAS in the military domain. Beyond the challenge of defining MHC, prominent issues in establishing human control are automation bias and system features such as self-learning abilities and software updates. While the definitions of RAS and AWS present a framework within which the systems exist, establishing human control is made difficult by the varying degrees of autonomy of systems themselves, which resembles a spectrum rather than a clear-cut categorization, as well as the different elements that make up a system. This paper defines autonomy itself and also adopts one of the commonly used classification frameworks for a system's degree of autonomy to avoid generalization of terms in the discussion.

Degrees of Autonomy

Determining the 'intelligence' of autonomous systems is complicated by the relational nature of intelligence and the human tendency to re-consider what computer models are perceived as such once a new generation of algorithms becomes achievable and operational. This reserves the status of 'Artificial Intelligence' for systems humans have yet to develop.[31] This affects how humans perceive AI-driven systems and how comfortable they are with using them, particularly in the military context, and, as such, which systems nation-states are comfortable with rolling out within their military forces. As new systems enter general use in civil contexts, humans become accustomed to them.[32] The result is continuous shifting of the ethical boundaries that dictate what is considered acceptable applications for and use of systems. Autonomy in the context of this discussion is defined as the following:

[30] Decommissioning is particularly relevant as the equipment can be sold to another military, whereby the risks discussed in this paper are still present, but offloaded to a third party, suggesting considerations that need to be introduced within arms control regimes.

[31] Scharre, *Army of None: Autonomous Weapons and the Future of War*, 242.

[32] Clarke, *Profiles of the Future: an Inquiry into the Limits of the Possible*.

Autonomy

Autonomy is the level of independence that humans grant a system to execute a given task. It is the condition or quality of being self-governing to achieve an assigned task based on the system's own situational awareness (integrated sensing, perceiving, analyzing), planning, and decision-making. Autonomy refers to a spectrum of automation in which independent decision-making can be tailored for a specific mission, level of risk, and degree of human-machine teaming.[33]

A further distinction is drawn between automatic, automated and autonomous systems.[34] Automatic systems are simple and threshold-based, whereby their action following a sensory response is linear, immediate and highly predictable. Automated systems are more complex and consider a range of inputs and variables before acting. Autonomous systems are goal-oriented and self-directed, meaning the operator may not understand the computational process that the system used to arrive at its conclusion (see Figure 3, p. 16). The distinction between automated and autonomous is difficult, as many existing RAS and AWS are forms of sophisticated automation, rather than actual autonomy.

An important difference to highlight is between the two distinctive meanings of autonomy in this context. The first refers to the degree of independent 'thought' or direction of action as well as the ability a system has to complete goals through computation not understandable by humans. This is often referred to as Artificial General Intelligence (AGI).[35] The second meaning of autonomy refers to the freedom of action granted to systems by humans, where a system is enabled to operate independently but is bounded by a strict set of rules, such as an automatic or automated system operating with minimal or no human oversight.[36] Most AI experts concur that AGI has not yet been achieved and is not set to be for the coming years, while complex automated systems that often operate autonomously, such as the Aegis Combat System, have been in military use for over forty years.[37] Therefore, it is crucial to distinguish between autonomy which grants machines freedom of 'thought', or at least determining a course of action based on their own computation, and autonomy which grants machines freedom of operation based on a set of rules that direct their operation.

[33] The definition is borrowed in full from Feickert et al., "U.S. Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress."

[34] Scharre, *Army of None: Autonomous Weapons and the Future of War*, 30–31.

[35] JASON, The MITRE Corporation, "Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD," 4.

[36] JASON, The MITRE Corporation, 4.

[37] Scott et al., "Modeling Artificial Intelligence and Exploring Its Impact."

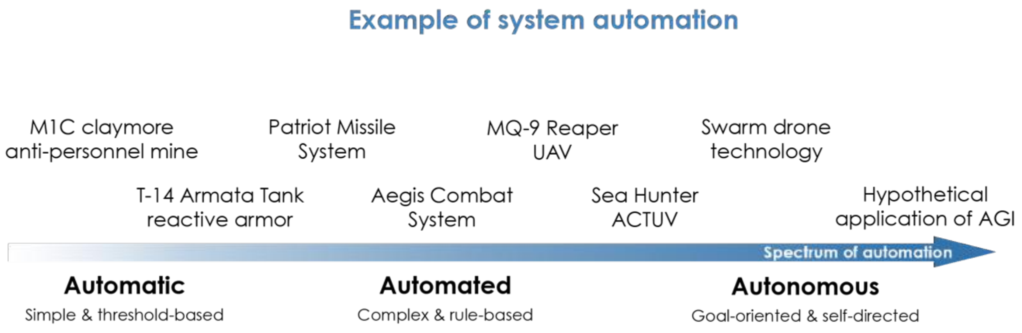

Figure 3. Examples of systems on the spectrum of automation[38]

Figure 3 illustrates how the degree of automation resembles a spectrum—rather than pre-defined categories—resulting primarily from technological advancements that are non-linear and manifest themselves in different machine functions. Certain sub-system functions may have a higher degree of autonomy than others, making the system more intelligent as a whole, but not enough to reach the next category, hence why the MQ-9 Reaper UAV is not considered fully within the 'autonomous' category above, for example. While the automatic–automated–autonomous scale can be perceived as a spectrum, to avoid generalizations in the discussion on autonomy, the various degrees of autonomy are placed in select categories, with the following framework adhered to throughout this paper:[39]

Direct control	A human operator has complete control over the observe-orient-decide-act (OODA) loop of the machine. This includes unmanned systems that are controlled by an operator through a machine interface. An example of this is the remote-controlled RQ-11 Raven miniature UAV.[40]
Semi-autonomous	A human operator is involved in sections of the OODA loop. An example of this is loitering munitions such as the Hero-400EC Extended Range Loitering System that requires a human to pre-identify a target but is self-guided once launched.[41]
Supervised autonomous	A human operator supervises and, if necessary, intervenes in the functioning of the autonomous system, but the OODA loop can function independently as a whole. Defensive systems such as the

[38] Adapted from Scharre, *Army of None: Autonomous Weapons and the Future of War*, 31.

[39] "Unmanned Systems Integrated Roadmap FY 2011-2036," 46. For the purpose of this paper, the taxonomy on the degree of autonomy is adapted to the terminology used by the Netherlands Armed Forces.

[40] "RQ-11 Raven Unmanned Aerial Vehicle."

[41] "Hero-400EC Extended-Range Loitering System."

	US Aegis Combat System can effectively operate without human input after initiation, but human intervention is possible.[42]
Fully autonomous	No human is involved in the operation of the system, but most advanced militaries agree on at least a minimal requirement for a human to decide to start and shut down a system. These systems are not as widespread as those in the previous categories, but examples exist in both research and development (R&D) and actual use. A non-robotic example is found in cyber defense, with machine learning algorithms that learn to defend systems from new types of malicious software autonomously, in ways such as capture-the-flag games, where two intelligent systems compete to attack or defend a network or system.[43] Meanwhile, R&D of ground, naval and aerial drone swarm technology is underway in China, Russia and the US, among others, introducing the possibility of fully autonomous RAS systems that operate collectively and independently through machine-to-machine communication.[44] Since the wider-scale introduction of these systems is likely in the (near) future, this category is included to stimulate debate on the ethics of fully autonomous systems as well.

Perception of human control in the use of all abovementioned types of systems is often disconnected from the actual extent of control. A prevalent heuristic is the instrumentalist perspective, through which technologies are perceived as 'tools' that are directly at the disposal, and hence, under the control of human users.[45] This is based on an underlying assumption that humans maintain agency over these tools and their operation.[46] An example is the often-suggested 'kill switch', where the idea is that meaningful human control is attained simply by the operator having the chance to either go along with or change a system's suggested course of action based fully on its own computation. However, the new generation of technologies is increasingly complex and purposefully designed to outperform narrow human tasks. Human agency is undermined by the cognitive inability of humans to keep up with the pace of algorithmic calculations of systems that operate with a higher degree of autonomy and independence.[47] This has different implications for operators that make decisions

[42] "AEGIS Weapon System." It can be argued that the Aegis is a sophisticated variant of an automated system, rather than autonomous. However, due to its goal-oriented approach and multiple operating settings, it is often referred to as a supervised autonomous system.

[43] Han et al., "Reinforcement Learning for Autonomous Defence in Software-Defined Networking."

[44] Long, "China Releases Video of 56-Boat Drone Swarm near Hong Kong"; Chung, "OFFensive Swarm-Enabled Tactics."

[45] Schwarz, "The (Im)Possibility of Meaningful Human Control for Lethal Autonomous Weapon Systems."

[46] Schwarz.

[47] Schwarz, "Intelligent Weapons Systems and Meaningful Human Control: An Uneasy Alliance," 4.

based on machine outputs and those that override a system, should it present incorrect or undesired outcomes. In the latter case, the ability of a human to intervene depends on the speed of the machine's operation, information available to the operator, and the time delay between the human input and the system's response.[48] This illustrates that the complexity of systems often undermines the instrumentalist perspective and suggests that there is a need to understand the extent to which humans have agency over systems and how they will continue to do so in the future.

To address the complexity of systems, this paper proposes a three-part structure to identify human control in RAS. The approach establishes human control through the perspective of the system life cycle, the sub-system operational structure and the OODA loop, and is presented below.

Identifying Human Control in Robotic and Autonomous Systems

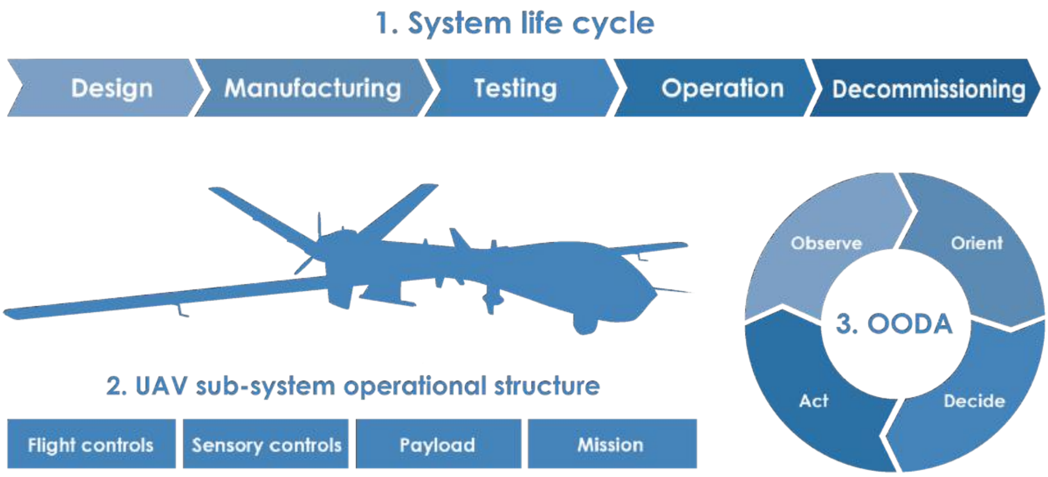

Figure 4. Various elements of human control in the operation of RAS

Figure 4 illustrates human control through three distinct perspectives, using a UAV as an example. As the diagram above demonstrates, human control can be identified throughout the (1) life cycle of a system, meaning that agreed upon degrees of control and/or oversight are maintained all throughout, or at least emphasized in sections such as testing and operation. A second, complementary approach, is to identify the controversial components at the (2) sub-system level. This separates functions such as movement control from payload in the degree of control and/or oversight required, and thus, provides a more nuanced view of requirements for meaningful human control in the overall operation of the system in question. A third, more specific part of the decision-making cycle is presented via the (3) OODA loop. The need for the

[48] Scharre, *Army of None: Autonomous Weapons and the Future of War*, 147.

degree of human control can be identified within the specific components of the decision-making loop of a system operating in a (semi-)autonomous mode.

1. Life Cycle of RAS

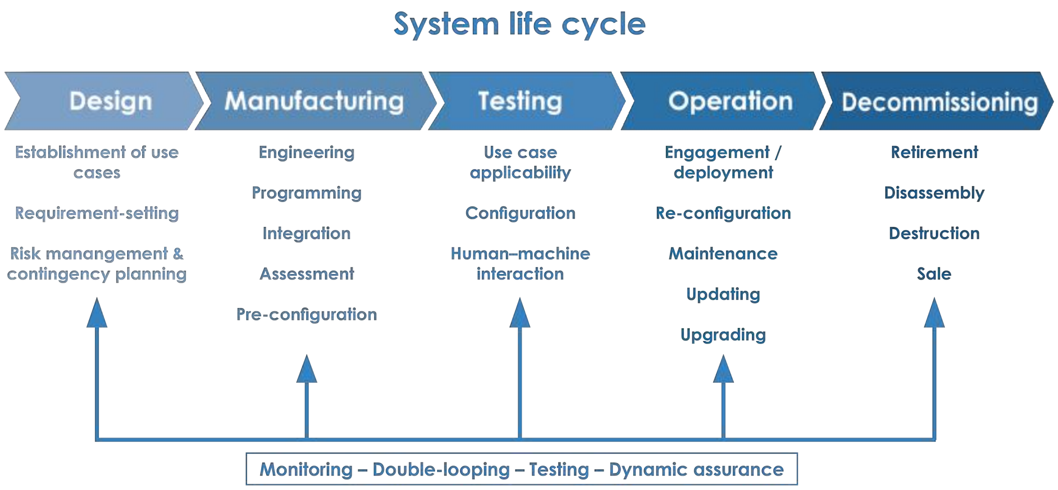

Figure 5. Life cycle perspective on human control

Design – The RAS life cycle perspective (Figure 5) seeks to contribute to the discussion on human control beyond the testing and operation sections of the system life cycle. This highlights the importance of establishing 'ethics by design', whereby ethical challenges are addressed early on during the design stage. The intention is to introduce control from the outset and to mitigate potential shortcomings in the operation of systems ahead of time.[49] Procurement of RAS may need to continue its shift from a traditional static tender process to a more dynamic, iterative process. This highlights the need for continuous monitoring, testing, and providing dynamic assurance of quality and functioning between all phases of the system's life cycle.[50] Through the involvement of end users, such as potential operators, during the stages of establishing use cases and requirement-setting, the appropriate user interface and user experience can be embedded into the design and manufacturing stages and improved as deemed necessary once tested and in use. This, in turn, aids how the users can work with, understand, and ultimately control the system.[51]

Manufacturing – In both the design and manufacturing stages (which are often outsourced to private contractors) military forces need to address the risks of being dependent on external commercial actors. At these stages, inadequate oversight when

[49] Floridi et al., "From What to How - An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices," 14.

[50] Arthur van der Wees, interview, 22nd August 2019.

[51] Henderson-Sellers and Edwards, "The Object-Oriented Systems Life Cycle," 144; Lehman, "Programs, Life Cycles, and Laws of Software Evolution," 1065.

working with contractors may result in not only the delivery of lower-quality equipment or systems not fully attuned to certain contexts, exacerbating the ethical risks of their operation. To address such possible shortcomings, there is a need to continuously evaluate the manufacturing process and review the 'factory settings' pre-configurations of the system. This way, RAS are tailored for the environments they will be deployed in, and through training and experience they can be reconfigured depending on specific issues or needs.

Testing – In testing, the people that will be working with the system in question need to gain sufficient knowledge of the system, have an understanding of its function, and be able to predict the response of the system to inputs in operational environments. At this stage, limitations in human–machine teaming can be addressed through continued monitoring and double-looping of human–machine interaction. Furthermore, use case applicability can be further refined in the testing stage and complemented with configuration for the identified use cases.

Operation – During the operation stage, system deployment and use should adhere to the principles of International Humanitarian Law (IHL) and be adequately adapted for use in a specific context. Static oversight is not a guarantee of meaningful human control. Such control requires experience-based system re-configuration, updating, upgrading, as well as the continuous monitoring of all implemented changes. Considering the increasing autonomy of systems, even routine activities such as maintenance[52] require ethical considerations: it cannot be left up to chance what the output of a system may be following a tweak or an update.[53] As such, there is a need for a dynamic risk assessment plan that spans the life cycle of RAS.

Decommissioning – The final stage in the life cycle is the decommissioning of RAS. There is a challenge brought about by the difference between hardware and software longevity. Furthermore, certain parts will be kept internally to be reused for maintenance purposes in other RAS units, but the primary ethical challenge at the decommissioning stage is the sale or re-use of RAS to others, be this in full or by parts.

[52] Boeing, "Maintenance Program Enhancements."

[53] European Commission, "Commission Regulation (EU) No 1321/2014 of 26 November 2014 on the Continuing Airworthiness of Aircraft and Aeronautical Products, Parts and Appliances, and on the Approval of Organisations and Personnel Involved in These Tasks."

2. Functional Complexity of RAS

Human agency should be recognized not only with respect to RAS itself, but also in regard to individual functions and components of RAS. Complex systems now feature a range of functions with varying degrees of autonomy. Figure 6 illustrates the composition of an average UAV, which has the following internal processes:

UAV Sub-system Operational Structure			
Flight Controls	**Sensory Controls**	**Payload**	**Mission**
Engine monitor	Strategic conflict detection & reaction	Actuators	Geographic Information System (GIS) database
Electrical monitor	Tactical conflict detection & reaction	Radar	Mission monitor
Virtual autopilot monitor	Visual/radar sensors	Image acquisition	Mission management
Flight plan	Awareness data fusion	Sensor data acquisition	Real-time data processing
Flight monitor	Long term planning		Scheduled communication
Air Traffic Control interaction	Traffic collision avoidance system (TCAS)		Storage module
Contingency management	Automatic dependent surveillance – broadcast		

Figure 6. UAV sub-system operational structure[54]

Functional complexity extends to systems beyond UAVs, where the parameters or functions may be different, but in similar fashion the range of sub-system autonomy may vary. Moreover, the number of software and hardware (sub-)components raises the question of compatibility in the long run when elements of software may be phased out earlier than other components, possibly rendering the remaining elements inoperable or reacting differently to the changing software. The other possibility is that the software, which receives continued updates, may outlive the electronic hardware. While throughout this paper no distinction is made between RAS as a whole and individual RAS functions, the issue is important to note when gauging the degree of autonomy of and human control over a system.

[54] Pastor et al., "An Open Architecture for the Integration of UAV Civil Applications, Aerial Vehicles."

3. OODA Loop in RAS

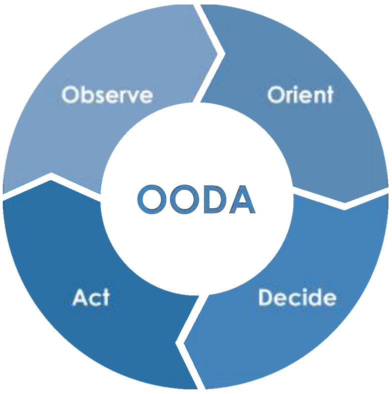

Figure 7. OODA loop

The final perspective through which this paper views human control within RAS is the observe-orient-decide-act decision-making loop.[55] Success in a military context is derived from the ability to shorten the OODA loop more quickly than an adversary, meaning decisions are carried out at a higher pace. With the advent of automation, OODA loops have continued to shorten, in many cases shifting the role of the human from system operator to supervisor, as computational speeds exceed the speed of human cognition. Advanced military forces expect OODA loops will shorten to fractions of seconds, meaning that ethical principles need to be established prior to deployment and included in the system design.[56]

An example of the changing human role in the OODA cycle is the comparison between a semi-autonomous weapon system and a supervised autonomous system. In the case of the former, the system carries out the appropriate calculations (i.e., observes and orients) while the decision to engage a target is retained by the human. In supervised systems, the human oversees the decision cycle and can override the system's decision-making process, but is otherwise not involved in the system's OODA loop. While humans are still involved in the OODA loops of semi-autonomous systems now, the advent of technologies such as drone swarming challenges the current understanding of human involvement and necessitates an understanding of how meaningful human control is retained in split-second decision-making loops.

4. Summary of the Three Perspectives

The three perspectives above together present a nuanced way of establishing meaningful human control in the use of RAS. It enables the development of guidelines based on the most challenging sections of the cycles, such as 'decide' in the OODA or the 'payload' in the sub-system structure. This enables the armed forces to continue expediting the OODA loop in sections with less ethical concern, while prioritizing the determination of human control in controversial sections. Combined, the perspectives bring to the forefront elements often disregarded in the debate on ethics, such as the design of systems and their decommissioning.

[55] Boyd, "The Essence of Winning and Losing."

[56] Scharre, *Army of None: Autonomous Weapons and the Future of War*, 23–24.

Determining Meaningful Human Control

Identifying and establishing meaningful human control can aid the process of establishing responsibility and accountability under International Humanitarian Law (IHL) in the use of AWS, particularly in the selection and engagement of targets.[57] There is a divergence in the interpretation of MHC, in terms of both the degree of control that should be required, and where and by whom this control should be maintained within the operational chain of a (semi-)autonomous system.[58] The critical issue is that there is no universally accepted definition, as on a national level each state interprets MHC to best suit their needs, while at the international level norm-setting has been impeded by states that benefit from the lack of clarity.[59] This lack of an explicit definition will continue to hamper the determination of responsibility in the use of RAS. There is an established consensus that humans are inherently responsible for actions of machines, but with increasingly complex systems that erode, or are perceived to erode human agency, it is necessary to outline exact features that would establish that the human control is 'meaningful'.[60] There is a further need to establish "*who* should exercise meaningful human control over *what*."[61] The current static approach of looking solely at the operator's control of the system negates the distributed nature of control that is spread across many individuals in the military decision-making cycle.[62] This reinforces the suggestion provided by this paper to view the identification of human control through the life cycle, sub-system functionality and OODA loop perspectives. Basic principles for MHC within AWS that have been proposed by normative actors, namely the ICRC and the non-governmental organization Article 36, include:

- Conscious human decisions, and timely judgment and intervention;
- Sufficient and accurate information on the outcome sought, the weapon system used and the context of its use;
- Transparency, predictability and reliability of the system linked to its design features; and
- Accountability for the functioning of the weapon system to a certain standard, such as IHL.[63]

[57] "Killer Robots and the Concept of Meaningful Human Control;" for more on the principles of IHL, see chapter 4.

[58] Sometimes also termed 'appropriate levels of human judgment' or sufficient human control'. See "Statement by France and Germany"; Docherty, "Heed the Call"; Sharkey, "Saying 'No!' To Lethal Autonomous Targeting."

[59] "Killer Robots Fail Key Moral, Legal Test."

[60] Santoni de Sio and van den Hoven, "Meaningful Human Control over Autonomous Systems"; Schwarz, "The (Im)Possibility of Meaningful Human Control for Lethal Autonomous Weapon Systems."

[61] Ekelhof, "Autonomous Weapons."

[62] Ekelhof.

[63] Ekelhof.

A practical consideration is that the normative actors often "articulate an idealized version of human control divorced from the reality of warfare and the weapons that have long been considered acceptable in conducting it."[64] This reiterates the argument that (semi-)autonomous systems have been adopted by modern armed forces over four decades ago and their use has generated little controversy.[65]

In the absence of a universal definition, it is worth considering the interpretations of MHC by a number of key actors engaged in the deployment of RAS, namely the UK, the Netherlands, the US, Israel, China and Russia.[66] The UK, an important player in the development of autonomous systems, emphasizes that "UK weapons will always be under human control as an absolute guarantee of human oversight, authority and accountability."[67] At the same time, however, the UK has a narrower definition of RAS than most other states, defining an autonomous system as one that is

> capable of understanding higher level intent and direction. [...] It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous unmanned aircraft will be predictable, individual actions may not be.[68]

The high degree of autonomy required by the UK's RAS definition means that current semi-autonomous systems may be excluded from the requirement of MHC by being deemed less advanced and therefore fall outside of the boundaries of RAS. Combined with the acceptance of unpredictability of certain functions of unmanned aircraft, the UK's understanding of MHC is distant from the basic principles presented by the ICRC and the non-governmental organization Article 36.[69]

The Netherlands Advisory Council on International Affairs (AIV) and Advisory Committee on Issues of Public International Law (CAVV) established that MHC is retained in the case of "an autonomous weapon [...] deployed after human consideration of aspects such as target selection, weapon selection and implementation planning, including an assessment of potential collateral damage."[70] Moreover, "in such cases, humans make informed, conscious choices regarding the

[64] Scharre and Horowitz, "Meaningful Human Control in Weapon Systems: A Primer."

[65] Ekelhof, "Lifting the Fog of Targeting: 'Autonomous Weapons' and Human Control through the Lens of Military Targeting."

[66] Most are derived from country statements at the Convention on Certain Conventional Weapons (CCW) meetings on lethal autonomous weapon systems.

[67] Evans, "Too Early for a Ban: The U.S. and U.K. Positions on Lethal Autonomous Weapons Systems."

[68] Development, Concepts and Doctrine Centre, "Unmanned Aircraft Systems - Joint Doctrine Publication 0-30.2."

[69] The UK's definition of RAS has 'higher level' expectations, which are beyond the majority of systems operated at this time. As a result, the discussion on MHC for UK's understanding of RAS may only apply to the complex systems envisioned with a higher degree of autonomy and thus downplay the degree of independence of existing systems.

[70] Netherlands Advisory Council on International Affairs, "Autonomous Weapon Systems: The Need for Meaningful Human Control."

use of weapons, based on adequate information about the target, the weapon in question and the context in which it is to be deployed."[71] This approach largely reflects the baselines set out by the ICRC.

Meanwhile, the US and Israel utilize the term "appropriate human judgment" rather than MHC.[72] During the 2016 Convention on Certain Conventional Weapons (CCW),[73] Israel argued that appropriate human judgment is already "built into the development of weapons systems, including at the design, testing, and deployment phases, and thus requiring meaningful human control is unnecessary."[74] The US also actively considers the entire RAS life cycle, meaning that "systems will go through rigorous hardware and software verification and validation (V&V) and realistic system developmental and operational test and evaluation (T&E) [...]."[75] The US Department of Defense directive on autonomy in weapon systems requires "traceable feedback on system status", explainability and predictability features, and has safety considerations for the human–machine interface.[76] It is therefore evident that countries have strongly varying positions on MHC, but some leading actors in RAS development agree on the importance of establishing some degree of human control in the design and procurement processes, rather than just in training and operations.

At the 2016 CCW meeting, China stated that "[t]he mode of human involvement and the human role [...] requires a strict definition and cannot be replaced by such vague concepts as 'human judgment' or 'meaningful human control'."[77] Internationally, China maintains its position that controllability remains a priority for any (semi-) autonomous military technology. However, on the same day that China reiterated its support for the development of a binding protocol banning the use of fully autonomous weapons at the 2018 CCW meeting, the country's air force published a statement that clearly demonstrated China intends to develop such systems anyway.[78] China is likely to advocate for international agreements that leave its own view on MHC slightly ambiguous, while delegitimizing the moral position of actors such as the US that have opposed adopting new laws on this topic just yet.[79]

[71] Netherlands Advisory Council on International Affairs.

[72] "Killer Robots Fail Key Moral, Legal Test."

[73] Lewis, "AI and Autonomy in War: Understanding and Mitigating Risks"; "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."

[74] "Killer Robots Fail Key Moral, Legal Test."

[75] US Department of Defense, "Directive 3009.09."

[76] US Department of Defense.

[77] "The Position Paper Submitted by the Chinese Delegation to CCW 5th Review Conference."

[78] Kania, "China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems," April 17, 2018; Klare, "Autonomous Weapons Systems and the Laws of War"; Mohanty, "Lethal Autonomous Dragon."

[79] Kania, "China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems," April 17, 2018; Kania, "Battlefield Singularity."

Despite the official statements at the CCW meetings,[80] the Russian defense ministry is clear about its intention to develop autonomous weapon systems, with some arguing that "Russia seeks to completely automate the battlefield."[81] Similarly to China, the country is on track to develop swarm units, which are groupings of autonomous systems that are inherently difficult to maintain human control over once deployed. The announcement by weapons manufacturer Kalashnikov that it will develop "a series of autonomous weapons using neural networks trained to autonomously track targets and fire on them" and Degtyarev's development of the "suicide tank" match defense officials' enthusiasm for robotization and lack of interest in human control as a prerequisite—no matter the definition.[82]

Traceability, Explainability and Predictability

A fundamental aspect of maintaining MHC is the operator's understanding of the algorithmic process' parameters, the outcomes presented as a result of the computation, and the ability to explain the machine's path to conclusion after the fact. Explainability is a prerequisite to determining some degree of operator's responsibility for the actions of a system. One of the ethical criticisms of RAS, and AI in particular, is the current lack of algorithmic transparency. Algorithms such as neural networks suffer from opacity as they operate as 'black boxes', whereby the path taken by the algorithm to arrive at the conclusion is often not traceable.[83] Beyond the 'black box' effect, algorithmic opacity arises from technical illiteracy, whereby creators write poorly structured code, or programmers on the receiving end are unable to understand the creator's intent.

Opacity is reinforced by the complexity, self-learning capabilities and scale of algorithms, with systems such as the F-35 fighter jet and self-driving vehicles requiring 24 million and 100 million lines of code, respectively.[84] As a result, algorithmic activity may not be traced and hence, in the event of a malfunction, the cause of the failure will not be determined rapidly. The diminished understanding an operator has of such systems reduces their ability to predict and/or explain the system's reasoning process. This may undermine the control that the operator has over the outcomes and

[80] "Statement of the Head of the Russian Federation Delegation, Director of the Department for Nonproliferation and Arms Control of the Russian Ministry for Foreign Affairs V.Yermakov at the Meeting of the State-Parties of the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons on Item 7 of the Agenda 'General Exchange of Views', Geneva, November 21, 2018."

[81] Sharkey, "Killer Robots From Russia Without Love."

[82] As for the suicide tank, "Once launched it can navigate autonomously to a target in silent mode and then explode with a powerful force to destroy other tanks or entire buildings.". See Sharkey; Gilbert, "Russian Weapons Maker Kalashnikov Developing Killer AI Robots."

[83] Preece, "Asking 'Why' in AI: Explainability of Intelligent Systems – Perspectives and Challenges"; Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," 178–79.

[84] Burrell, "How the Machine 'Thinks'"; Mittelstadt et al., "The Ethics of Algorithms," 3–7; Scharre, *Army of None: Autonomous Weapons and the Future of War*, 157.

hence, the responsibility for its (mis)use.[85] However, given the sheer scale of code in complex systems, the expectation for the operator to understand granular functions, especially in the operation stage of the life cycle is unrealistic.

Progress is being made in understanding the internal workings of algorithms, by means such as the Local Interpretable Model-Agnostic Explanations, an algorithm that explains the prediction of any classifier algorithm in a method interpretable by humans.[86] An alternative is the use of "logic flow diagrams", which summarize sets of code and enable the operator or supervisor to trace the process via critical junctures such as the OODA loop steps, thus maintaining a macro perspective of the system performance.[87] This approach is comparable to the use of a vehicle without explicit understanding of its mechanical functions but with knowledge of the error signals displayed on the driver dashboard and their meaning.

Algorithmic systems predominantly operate based on historical training data to make future assessments and predictions. The need for quantification renders contextual non-numerical data—such as an individual's behavior and body language that are observed rather than measured—potentially invalid in the algorithmic decision-making process. This means elements that cannot be easily quantified are likely to be excluded from the calculation.[88] The result is a system that can function successfully in a controlled environment with defined parameters, but in a real-world scenario, where parameters are less defined, algorithms present outcomes based on a biased set of numerical inputs.[89] Moreover, the algorithms are often "embedded at the backend of systems, [...] with no consumer-facing interface. Their operations are mainly unknown, unseen, and with impacts that take enormous effort to detect."[90] In a high-intensity setting, this further establishes the need for extensive explainability, as the operator has to be acquainted with the system's intricate parameters and be familiar with its shortcomings between controlled and real-world scenarios.

When combined with other risks discussed in this chapter, the 'black box' effect can impede the functioning of human–machine teaming. Beyond understanding the reasoning process the system undertook to arrive at the conclusion, it is important for an operator to be able to anticipate how RAS will react in any given situation, particularly when it comes to a real-world scenario following controlled testing. A

[85] Preece, "Asking 'Why' in AI: Explainability of Intelligent Systems – Perspectives and Challenges"; Mittelstadt et al., "The Ethics of Algorithms," 5, 10–12.

[86] Ribeiro, Singh, and Guestrin, ""Why Should I Trust You?"

[87] Wu et al., "Research and Application of Code Automatic Generation Algorithm Based on Structured Flowchart"; Kumar et al., "Algorithms, Flowcharts, Data Types and Pseudocode."

[88] Schwarz, "The (Im)Possibility of Meaningful Human Control for Lethal Autonomous Weapon Systems."

[89] Schwarz, "Intelligent Weapons Systems and Meaningful Human Control: An Uneasy Alliance," 11.

[90] Crawford and Whittaker, "Artificial Intelligence Is Hard to See."

system may be predictable in a controlled environment for a programmer but may not maintain the same properties in a real-world scenario for an operator, supervisor and/or commander. Moreover, predictability of actions does not guarantee predictability of outcomes, which once again, can be influenced by the operational environment.[91] The discrepancy between training and combat, which may result in a predictability gap, is discussed further in Chapter 3.2.

Self-learning Abilities and Software Updates

The evolutionary nature of algorithm-driven systems, both as a result of self-learning properties and software updates, has the potential to considerably affect explainability of systems' actions. Self-learning AI that independently develops its understanding of the surrounding environment may limit human control over the system's operation. This is underpinned by the exponential growth of the capabilities of machine learning (ML) algorithms, such as, among others, neural networks and reinforcement learning.[92] As systems become increasingly complex, "in a steady progression the programmer role changes from coder to creator of software organisms."[93] Programmers are transitioning from maintaining control over the software code, to setting the algorithmic parameters and, depending on the algorithm, the network architecture for the algorithm to operate within. Reinforcement learning algorithms are particularly challenging, as they are designed to learn from their immediate environment.[94] The result of this is that next-generation algorithms no longer operate on pre-determined rules and can change their functionality, meaning humans often cannot understand the calculation made to arrive at the conclusion.

The evolving functionality is compounded by the involvement of third parties, often private companies, which are responsible for designing and supplying the system. This illustrates how human control is affected in the aforementioned life cycle and OODA perspectives. A further layer of complexity is added by the distancing of the programmer from the system, meaning the response time to faults and malfunctions is increased.[95] While the private contractors who design and manufacture the systems are now often deployed alongside the military, the potential inability of their military counterparts to understand RAS undermines meaningful control in the system's use, as it is the military operator who makes substantive decisions over the system's

[91] United Nations Institute for Disarmament Research (UNIDIR, "The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward."

[92] Scott et al., "Modeling Artificial Intelligence and Exploring Its Impact."

[93] Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata."

[94] An example of this is the AlphaGo Zero algorithm that mastered the game Go without making use of historical training datasets based on human inputs. United Nations Institute for Disarmament Research (UNIDIR, "The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence."

[95] Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata."

utilization. Depending on one's interpretation of MHC, there is potential for a loss of human control over time. This is a result of not only increasing sophistication in the AI that powers the RAS, but also the subsequent software updates that may reduce the military operators' understanding of the system over time.

This issue arose in at least two of the four scenarios in the HCSS expert session. Participants argued that operators would need training with the machine following each subsequent software update, as it could alter the machine's behavior and, as a result, the predictability of its outputs.[96] In reference to scenario 3, one participant noted the difficulty brought about by software updates, which meant that personnel had to get used to the change in the machine's method of operation, but the update could also unwittingly alter other capabilities. This begged the question of how much training is reasonable or necessary under these circumstances to ensure sufficient (re)familiarization with the updated system. Therefore, the RNLA would need to determine whether retraining is necessary for software updates involving all components of RAS or only for specific, pre-defined components.

To highlight the compounding effect of the aforementioned factors, an interesting example from civil aviation is that of the two Boeing 737 MAX 8 crashes, and the subsequent grounding of all MAX 8 aircraft worldwide in the Spring of 2019.[97] While the case is outside the military domain, it involves Boeing, a manufacturer which doubles as a defense contractor and highlights the challenges of relying on external contractors, as well as continuous software updating and system modification. In a cost-cutting bid, Boeing supplemented major mechanical engine re-design in the 737 MAX 8 with sensors and an additional automated system. The system aimed to compensate for the change in engine design and its new position under the wing.[98] The pilot manual included with the new aircraft did not sufficiently inform pilots familiar with other Boeing 737 variants of the changes introduced in the MAX 8.[99] This meant the pilots were not well trained with the machine interface, and as has been suggested following the incidents, led to the inability of pilots to override the system

[96] In scenario 2 (testbed), experts argued that the system would have to undergo testing after each software update, to ensure that the operator has up-to-date understanding of the system's behavior.

[97] Based on preliminary findings, the cause of the two crashes (Lion Air Flight 610 & Ethiopian Airlines Flight 302) involving the Boeing 737 MAX 8 has been a malfunction of the Maneuvering Characteristics Augmentation System (MCAS). The system was introduced to offset the engine design changes made to the 737 MAX 8 from the previous models in the 737 series. More specifically, the change was the relocation of the engine position under the wings, resulting in a shift of the centre of gravity of the aircraft. To control for this, Boeing installed the MCAS, which was supposed to use sensors to indicate to a computer if the aircraft was stalling mid-air. However, the system also reacted in cases where the sensor input was contradictory. In both incidents, the system falsely identified the aircraft's angle-of-attack as excessively high and sent the aircraft into a nose dive as to prevent it from stalling midair. The aircraft were not in fact stalling, and the MCAS misidentifications and automatic reaction, combined with the planes' close proximity to the ground (as both incidents occurred at take-off) left the pilots no time to override the MCAS and take manual control of the aircraft. Travis, "How the Boeing 737 Max Disaster Looks to a Software Developer"; Lu et al., "From 8,600 Flights to Zero: Grounding the Boeing 737 Max 8."

[98] Travis, "How the Boeing 737 Max Disaster Looks to a Software Developer."

[99] Hawkins, "Everything You Need to Know about the Boeing 737 Max Airplane Crashes."

in a short span of time when it malfunctioned. The operation of a highly automated system—the 737 MAX 8—was hampered by substantial functionality alterations that were not well communicated to its users. Boeing presented the aircraft as closely related to previous 737 models, thus suggesting that extensive re-training was not necessary, in turn resulting in incomplete preparation of the pilots to use the new aircraft. The delivery of the unsafe 737 MAX 8's highlights the additional risk of poor oversight of private contractors, an issue discussed later in this chapter.[100]

The legal cases following the incidents are on-going, but the issue of fragmented responsibility is already evident. Who is responsible for the incidents? At the macro level, is it the airline companies, Boeing, pilot training organizations and flight simulator operators, regulators of airlines in the host countries or the regulators of Boeing's host country (in this case the US Federal Aviation Authority (FAA))? At the micro level, is it the pilots, airline engineers, Boeing engineers, Boeing software developers accountable or the FAA inspectors? The issue of attributing responsibility is further complicated by public–private and cross-border divides, whereby accountability can be hampered by national laws and (international) contracts. While this case is outside the military context, it clearly demonstrates the discussed issues unfolding in a real-life context and reflects the issue of distributed responsibility in the military decision-making cycle, which also involves many actors, from programmers and manufacturers, to operators and commanders. Distributed responsibility is discussed further from the legal perspective in Chapter 5.

3.2 Human–Machine Teaming

The relationship between the military personnel and operators, and the machines they work with, is another important aspect of human agency. How RAS are developed and deployed will deeply impact the people working with them.[101] Aside from determining the level of MHC and being able to understand and explain the reasoning process of the system, human–machine interaction encompasses several other issues, raised both in the literature and the expert session. These are automation bias, interoperability challenges, trust in the manufacturer of the system versus knowledge of the system itself, comparability of testing environments to real scenarios and the anthropomorphizing of machines.

[100] Stewart, "The Boeing 737 Max 8 Crashes and Controversy, Explained."

[101] Roff and Danks, "'Trust but Verify.'"

Automation Bias and Complacency

Human overreliance on and uncritical trust in computer-based decision-making, otherwise known as 'automation bias', impacts the level of control that operators can exercise over RAS.[102] This is a human tendency to ignore or not seek out contradictory information to the outputs of an automated process, due to a perception of machines' superiority in accuracy.[103] 'Automation complacency' is not dissimilar, but while automation bias refers to excessive trust in a system, automation complacency concerns substandard attention to and monitoring of a system's output, on the assumption that the output is reliable.[104] Both bias and complacency lead to problems of process malfunction misidentification, anomalies and failure, as well as delays in the response time of human intervention resulting from insufficient oversight. The latter is no less critical than outright failures in oversight, as in fast-paced combat situations, an untimely response can have serious implications for the outcomes of the use of RAS and AWS.[105]

As a result of automation bias, meaningful human control is reduced. Humans may place a disproportionate amount of trust in the automated processes they are meant to control or supervise. An example of this is the shooting down of an Iranian passenger jet in the Persian Gulf by the United States Ship (USS) Vincennes in 1988.[106] Amidst an engagement between the USS Vincennes and Iranian forces, the automatic targeting-and-firing system Aegis misinterpreted the passenger jet for a military fighter jet and the naval vessel crew shot down the civilian aircraft. This highlights a failure to recognize and challenge shortcomings in a computerized decision-making process, particularly in a high-intensity combat situation when what appears to be an imminent threat can lead to a lethal counterattack. While the error occurred due to an incorrectly pre-selected operation setting, the automation of the weapon system in this instance significantly reduced the time for human decision-making and intervention. The operators' lack of due diligence in this case highlights the risks of the coupling of narrowing response time frames with automation bias and/or automation complacency.

Consideration of automation bias and complacency is therefore critical in establishing meaningful human control in the use of RAS, as "technologists tend to push to

[102] Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems."

[103] Cummings.

[104] Parasuraman and Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration," 382.

[105] NASA's Aviation Safety Reporting System (ASRS) defines complacency as "the state of self- satisfaction that is often coupled with unawareness of impending trouble," see Bhana, "By the Book - Good Written Guidance and Procedures Reduce Pilots' Automation Complacency"; Parasuraman and Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration."

[106] Kania, "The Critical Human Element in the Machine Age of Warfare."

automate tasks as fully as possible."[107] There is a need for explicit understanding of how increased process automation affects human cognition, and in turn human–machine teaming.[108] Determining the balance between the benefits of automation and its risks is particularly important in a military context, where bias and complacency can reduce explainability and, in turn, the responsibility of operators. The RNLA should seek to understand at what degree of autonomy the limitations of human cognition in the oversight of RAS offset the benefit of increased autonomy. An alternative function is one similar to operator assist, whereby the systems enhance human functions, rather than replacing them entirely.[109] In this instance a critical consideration is testing and validation of human–machine interfaces that diminish the effects of automation bias and complacency.

Trust or Knowledge

Another point of discussion is that RAS—whether as logistics, weapons, or otherwise—tend to be developed in order to enhance militaries' all-round capabilities.[110] This justification assumes both that the RAS will function as intended, and that the users and/or operators trust the systems to the extent that these can function as intended.[111] This issue is particularly potent when considering the role of third-party contractors in supplying RAS, particularly AWS. There is a risk of the military outsourcing excessive control to the contractor and not being fully informed on the reasoning process of RAS, particularly when contractors are involved in the operation of the system. This relates to the case involving the Boeing 737 MAX 8, which supplied hundreds of aircraft to over 40 airlines worldwide before the fatal issue was identified.[112] This requires a due-diligence process before and throughout the engagement with external contractors. Oversight has to be maintained in the design of the systems and throughout their deployment, since as long as they continue to function, they will depend on externally developed software and hardware to do so. Therefore, the RNLA need to ensure that it has thorough knowledge of the functional parameters of the (semi-)autonomous systems they purchase, rather than basing the acquisition and use on the trust placed in the contractor and the operator of the system.

[107] Miller and Parasuraman, "Designing for Flexible Interaction Between Humans and Automation," 58.

[108] Hoijtink and Leese, *Technology and Agency in International Relations*, 50–53.

[109] JASON and The MITRE Corporation, "Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD," 54.

[110] Roff and Danks, "'Trust but Verify.'"

[111] Development, Concepts and Doctrine Centre, *Human-Machine Teaming*.

[112] Lu et al., "From 8,600 Flights to Zero: Grounding the Boeing 737 Max 8."

Interoperability with Partner Forces

Figure 8. US 82nd Brigade Engineer Battalion launches a Puma aviation system during a multinational joint equipment training brief in Germany, April 2018. Photo: Spc. Dustin D. Biven/US DOD

A variation of the 'trust or knowledge' issue was identified during the expert session. It concerns how the military can engage in combat alongside technologically advanced allied military forces operating RAS that the former has not yet worked with.[113] In the expert session discussions, participants concluded that for the joint use of an autonomous system in the military, it is crucial that the troops using it fully understand the system they are using and can predict the behavior of RAS in the situation or environment that the system is used in. However, the necessary level of (training) experience with a system, or how much prior information on the system from a partner, remained a point of discussion, and is clearly a topic that requires further research.

Opinions on the level of training ranged from troops needing to be well acquainted with and have trained with the system for years on the one hand, to troops in collaboration with trusted partners receiving a certain amount of information on the systems used, including testing reports, relevant indicators and embedded rules of engagement (ROE). Difference in levels of not only confidence, but also trust in fellow soldiers as opposed to RAS, partly comes down to the extent to which one can understand the reasoning process of each. Even if a human acts irrationally, there is a certain reasoning process behind the actions that can be explained and likely

[113] In scenario 3, Dutch soldiers, hypothetically, are faced with a choice between deploying tired Dutch soldiers to secure a facility, or deploying Danish-made and -operated LAWS, the exact parameters and reasoning process of which are unknown to the Dutch contingent.

understood afterward. Therefore, a level of understanding of the reasoning process that guides the actions of RAS is needed and this should likely be communicated to partner forces that are collectively engaged in a single operational theatre. This issue highlights the need for standardization frameworks between allied forces that deploy and operate RAS during joint missions. This can further be extended to incorporate interoperability with private military contractors that are deployed alongside many advanced militaries in operational theatres today.

Testing Environments

Most RAS are tested in safe and controlled environments, so the response of the systems and the operators in a real, high-intensity and uncontrolled scenario is often unknown. For the operator, it is important to determine how they will respond to the machine's outputs, particularly when the intervention or decision-making timeframe is narrowed by the rapid computational processes. An example of where this becomes problematic is the 2004 friendly fire incident involving the downing of two British and US fighter jets by a US Patriot system over Iraq, resulting in the deaths of three aircrew.[114] The lack of training of the operators and their unfamiliarity with the machine interface resulted in the inability of the operators to intervene in the Patriot's decision to fire at allied aircraft.[115] As a result, the lack of training with the system and insufficient familiarity with the reasoning process of the Patriot in this case, further reinforced the operator's dependency on the conclusions of the system's internal processes.

From the perspective of RAS, it is unclear how a system, particularly with self-learning abilities, will respond to situational uncertainty and nuance otherwise not present in controlled testing environments, particularly if the system depends on machine learning algorithms. The issue was discussed in reaction to scenario 1 (in Appendix B) in the expert session, where experts doubted the ability of the AWS to distinguish its high-value target from other militants and civilians in a poorly lit cave system. There is an evident need to simulate highly realistic combat scenarios and even test equipment outside of controlled environments. However controversial, Russia has used its recent Syria campaign to test various autonomous systems in combat, namely an underwater unmanned system and an electronic warfare unit.[116] The RNLA may therefore consider studying the testing approaches of other states and determine how they can best emulate it without violating ethical principles.

[114] Cummings, "Automation Bias in Intelligent Time Critical Decision Support Systems."

[115] Cummings.

[116] Grishenko, "Российский Подводный Робот Выполнил Боевую Задачу в Сирии"; Bendett, "In AI, Russia Is Hustling to Catch Up."

Anthropomorphizing the machine

The final point of discussion is that individuals interacting with artificial systems tend to anthropomorphize (humanize) them, whereby they "attribute minds to computers and perceive robots as agents."[117] When RAS are "seen as more than just a tool to achieve an effect," this may hinder the intended functions.[118] This highlights the human need to attribute features to inanimate objects and paradoxically relates to the instrumentalist perspective introduced earlier in Chapter 3.1.[119] The "efforts at building self-explicating machines in their more sophisticated forms now adopt the metaphor of the machine as an expert and the user as a novice or student,"[120] demonstrating that humans are transitioning from perceiving machines as tools under their control (instrumentalist perspective), to humanizing them as they increasingly replicate human functions. This is a result of design that deliberately seeks to develop systems that exceed our control and cognitive abilities.[121] The limitations that result from humanizing machines are primarily manifested in two ways. The first is the operator becoming attached to the machine, which defeats the purpose of having RAS replace soldiers in combat and influences the operator's risk perception of the situation.[122] The second is the operator humanizing the machine and anticipating a human way of thinking, thus overseeing the limitations of algorithmic outputs.[123] The risk is that the operator is unable to develop a mental model to cope with the system and handle the system's failures, resulting in undesired effects within human–machine teaming.[124]

3.3 Summary

This chapter has highlighted the difficulty in identifying the extent of autonomy as well as determining meaningful human control and establishing it in practice. While ethicists debate practitioners on whether autonomous weapons should be banned, less controversial applications of RAS will continue to permeate the military domain. Establishing meaningful human control should be preceded by the identification of the type of autonomy displayed in a system, within which sub-system functions its

[117] Verdiesen, "Agency Perception and Moral Values Related to Autonomous Weapons," 96; Schwarz, "Intelligent Weapons Systems and Meaningful Human Control: An Uneasy Alliance," 4,11.

[118] Krishman, *Killer Robots: Legality and Ethicality of Autonomous Weapons*; Hsu, "Real Soldiers Love Their Robot Brethren"; Schwarz, "The (Im)Possibility of Meaningful Human Control for Lethal Autonomous Weapon Systems"; Robert, "The Growing Problem of Humanizing Robots."

[119] Sharkey, "The Evitability of Autonomous Robot Warfare."

[120] Suchman, *Human-Machine Reconfigurations: Plans and Situated Actions*.

[121] Gunkel, "Other Things: AI, Robots and Society," 60; Schwarz, "Intelligent Weapons Systems and Meaningful Human Control: An Uneasy Alliance," 4, 11.

[122] Giger et al., "Humanization of Robots."

[123] Robert, "The Growing Problem of Humanizing Robots."

[124] Brenton and Bosse, "The Cognitive Costs and Benefits of Automation."

present, and how this affects the OODA loop of RAS. As the armed forces seek to harness efficiency gains of AI and maintain their competitive edge, research and deployment of RAS is likely to persevere.

With countries seeking to shorten the OODA loop, there is an opportunity in distinguishing between ethically controversial and undisputed functions of RAS and using this to streamline automation. The former, primarily concerning elements of weapons delivery, require further consideration and addressing of ethical questions, while the latter, such as movement or sensory control, can continue to be automated. To confidently pronounce that an operator/supervisor has meaningful control at the 'decide' stage of the OODA loop requires the assertion that human control was maintained in the design and testing of the system as well as within its sub-system functions. This, for example, involves understanding of sensory outputs that inform the 'observe' and 'orient' sections of the OODA loop, before a decision can be made. Following this, establishing MHC requires an understanding of where, how and by whom it should be maintained. Finally, challenges around maintaining MHC, among which are cognitive limitations in human–machine interaction, shortcomings in machine development, and risks of procurement from third parties, must be addressed to ensure that responsibility for the deployment and the use of RAS can be established, and that those responsible can be held accountable.

4 | Human Dignity

International Humanitarian Law (IHL) has a strong footing in ethics, and its basis lies in the attempt to attain and maintain respect for the dignity of human life in the disarray of war. To be exact, IHL refers to international rules intended to protect people and property that are, or may be, affected by (inter)national conflict through setting limits on how conflicting parties may choose their methods and means of warfare.[125] There are currently no specific bans or regulations under international law that make RAS unlawful per se. At the same time, however, the use of RAS in conflict still means that those deploying RAS are bound by the obligations of IHL.[126] IHL principles are therefore an important part of addressing ethical issues pertaining to military RAS in general, and the use of force when deploying AWS.

Currently there are no offensive (supervised) AWS ready for deployment that in offensive situations could satisfy IHL obligations.[127] This makes for a relatively easy conclusion on non-deployment at the present stage.[128] However, it was highlighted in the expert discussions that once the technical ability to comply with said IHL obligations is available, the difficulty will be assessing under what circumstances the use of AWS could be permissible, and what the society in question deems crucial to maintaining some level of humanity on the battlefield.

This chapter on human dignity is split into two sub-chapters, aimed at addressing the ethical discussions outlined in the previous two paragraphs. In Chapter 4.1 the key principles that govern hostilities are laid out and structured along the lines of International Humanitarian Law.[129] This sub-chapter also addresses the ways in which these principles may affect or be affected by increased integration of RAS into the armed forces. Chapter, 4.2, describes the main arguments that feature in debates on how increased use of military RAS may affect the status of and respect for human dignity. Two main points of contention were identified before the expert session for the topic of human dignity: whether decisions that have always been inherently human could and should be substituted with computer processes—especially if they involve life-and-death situations; and whether there may be a point in time, or a particular situation, where substituting certain human tasks or operations with RAS may be considered more ethical, rather than less. These are two central questions that

[125] Bouvier, "International Humanitarian Law and the Law of Armed Conflict," 13.

[126] Davison, "Autonomous Weapon Systems."

[127] Boothby, *New Technologies and the Law in War and Peace*; Chehtman, "New Technologies Symposium."

[128] Boothby, *New Technologies and the Law in War and Peace*.

[129] In military spheres preference is sometimes given to the phrase 'law of armed conflict' (LOAC) rather than International Humanitarian Law. However, the authors will refer throughout this paper to IHL as it is more widely used. See e.g. Bouvier, "International Humanitarian Law and the Law of Armed Conflict," 13.

feature in the discussion of how RAS may enhance or hamper respect for human dignity.

4.1 Ethics and International Humanitarian Law

This section lays out the key principles of IHL applicable in armed conflict. These are proportionality, military necessity, distinction, and, as a more general guiding principle underpinning the conception of ethics, humanity.[130] Throughout each subsection there will be an explanation of what each principle entails in general, as well as how it affects or is affected by RAS—and by AWS in particular.

Proportionality

The first principle of IHL dictates that actions should always be proportionate. 'Proportionate' in this context means that expected incidental harm to civilians or civilian objects—also known as 'collateral damage'—should not be excessive in relation to the concrete and direct military advantage anticipated.[131] The standard by which this is assessed is that of a "reasonable commander or combatant who weighs the expected collateral damage against the anticipated military advantage in good faith, based on information available at the time of the attack."[132] Whether this standard suffices when it comes to using RAS or how exactly it would apply, makes the breakdown of RAS and human control illustrated in Figure 4 in Chapter 3.1 a useful guideline.

The degree of leeway offered by the proportionality principle as described in IHL has often been interpreted differently, and the interpretation has also shifted over time. Proportionality is viewed both as a permissive and a restrictive principle. On the one hand, the fact that states are themselves responsible for weighing military advantage from certain actions against the degree of civilian damage, could be viewed as permissive. On the other hand, the principle could be restrictive in that it may hamper military objectives, as more scrutiny is aimed at the justification behind individual attacks. With a shift from interstate conflict to more asymmetric forms of warfare, there tends not to be a clear-cut start or finish to a conflict, and properly establishing 'military advantage' can be difficult even for experienced military commanders.

[130] Within IHL the principle of humanity, also referred to as the Martens Clause, is often discussed in more straightforward terms as meaning the prevention of unnecessary suffering.

[131] Additional Protocol I, Article 51(5)(b) concerning the conduct of hostilities prohibits attacks when the civilian harm would be "excessive in relation to the concrete and direct military advantage anticipated." See Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I).

[132] Netherlands Advisory Council on International Affairs, "Autonomous Weapon Systems: The Need for Meaningful Human Control," 24.

Therefore, there is a persisting challenge of ensuring clear compliance with the proportionality principle by an autonomous system, be it defensive of offensive.

For AWS in particular, a difference can be made between the so-called 'easy proportionality problem' and 'hard proportionality problem'. The former concerns how to minimize collateral damage by using the most appropriate weapon and target, or in other words, taking all necessary precautions to minimize the damage done to civilians and civilian objects.[133] The 'hard proportionality problem' concerns the decision on whether or not to use force in the first place. This decision depends on how a commander weighs the balance between civilian lives and the wider military goal of the mission.[134] The 'hard' problem therefore concerns contextual factors beyond the specific situation at hand. Machines may be(come) better than humans at quickly assessing quantitative, computable elements of an attack, such as blast effect or number of potential civilian casualties. However, qualitative elements like the direct and indirect military advantage versus the civilian damage done remains, and may continue to remain for the foreseeable future, in better hands with humans.[135] The appreciation and weighing of a certain attack within the complicated context of a mission's larger military strategic aims, as opposed to within only the attack itself, involves difficult and occasionally morality-heavy decisions for commanders, and remains therefore a point of concern for the deployment of RAS.[136]

From the HCSS expert session discussions, it appears that people hold RAS to higher standards than humans when it comes to accepting mistakes on the battlefield. In one of the expert session scenarios, an AWS had an accuracy rate of 99.95%, but the 0.05% was already viewed as a major issue in light of civilians present in the operational environment.[137] Most participants remained undecided on the deployment of AWS in this particular scenario. On the one hand, there were major concerns over the machine's ability to fulfill the requirements of proportionality. It was also acknowledged, however, that the use of RAS in the scenario in question would—in the case that no mistakes occurred—be the most effective and least risky option for the troops, given the almost impossible circumstances of the scenario.

[133] Ekelhof, "Autonome Wapensystemen: Wat We Moeten Weten over de Toepassing van Het Humanitair Oorlogsrecht En de Menselijke Rol in Militaire Besluitvorming," 198.

[134] See Sharkey, "The Evitability of Autonomous Robot Warfare," 789.

[135] Ekelhof, "Autonome Wapensystemen: Wat We Moeten Weten over de Toepassing van Het Humanitair Oorlogsrecht En de Menselijke Rol in Militaire Besluitvorming," 198.

[136] van den Boogaard, "Proportionality and Autonomous Weapons Systems"; Sparrow, "Building a Better Warbot."

[137] See scenario 1 'Killerbot' in Appendix A for the full description of the situation.

Military Necessity

The second principle of IHL is the principle of military necessity. Following this principle, the use of RAS may only result in the use of force for legitimate military objectives, and every injury caused—even against enemy combatants—is only excusable in as far as it was absolutely necessary.

When it comes to targeting,[138] the necessity principle in practice often encompasses two questions that must still be answered after other obligations under IHL are shown to have been complied with:

a) Is the action required for direct military advantage, or, as the US Air Force puts it, "required to quickly and efficiently defeat the enemy?"[139]

b) Is the target of the action a valid 'military objective'?[140]

If question b were to be answered with a 'no', the principle of distinction (which will be explained in the next section) would per definition be violated, making the action impermissible under IHL. In judging the extent to which certain military functions could be performed by RAS, especially in the case of (semi-)autonomous weapons, it is therefore crucial that the system in question possesses the ability to distinguish valid military targets (see question b). This is often a context-heavy question—and once it is answered, the even more environment-dependent issue of direct military advantage will likely remain. This will require forms of human–machine teaming at least in the near future, with humans' experience, ability to draw from varying contexts, and creativity remaining relevant in decisions on the use of force.

Even more so than proportionality, the principle of military necessity remains controversial. Military necessity within IHL recognizes (gaining significant advantages in) winning a war as a legitimate consideration towards the use of force and legitimizes collateral damage to an extent. The principle can be seen in both a more permissive and a more restrictive light. A relatively widely accepted view falls in the middle, namely that the principle of necessity is itself, in general, a permissive

[138] The word 'targeting' does not refer to only the (kinetic) action against a target but rather, it indicates the larger military decision-making process. The various phases in this process are set out for example in Ekelhof, "Autonome Wapensystemen: Wat We Moeten Weten over de Toepassing van Het Humanitair Oorlogsrecht En de Menselijke Rol in Militaire Besluitvorming," 199–202; Ekelhof, "Lifting the Fog of Targeting: 'Autonomous Weapons' and Human Control through the Lens of Military Targeting."

[139] "Annex 3-60 – Targeting. Appendix A: Targeting and Legal Consideration. Basic Principles of the Law of War and Their Targeting Implications," 89.

[140] According to Article 52 of Additional Protocol I to the Geneva Convention, military objectives are "those objects which by their nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage."

principle, while the later-discussed principle of humanity has a necessary limiting function that counterbalances this.[141]

The issue of military necessity was raised frequently in scenario 4 of the expert session.[142] The experts weighed whether the deployment of a certain autonomous defensive system at the border that could identify and shoot down aircraft deemed to be a threat was necessary, due to the risk of two types of accidents that could occur. The first was the accidental hitting of non-military aircraft that violate the airspace, and the second was the shooting down of military aircraft that flew near the border, but rather than being a threat were actually only flying towards the system to test its parameters and the willingness of the defending state to engage the aircraft. One set of experts concluded that using the system was the most ethical approach to defense—if the intentions and the system's parameters had clearly been communicated to the adversary's higher command. This way, the defending force had explicitly delineated its position on the engagement of targets.

The obligation to ensure military necessity before resorting to the use of force is one that, in the context of AWS debates, turns back to the concept of human control. RAS can reduce the amount of time needed to come to conclusions or to make decisions—something that is becoming increasingly important both on the physical battlefield and in the cyber domain.[143] The way in which RAS can speed up decision-making is a double-edged sword, however. The more this speed surpasses what human reasoning is capable of, the more military technologies may shift from being largely diagnostic or descriptive to becoming more predictive or even prescriptive. There are two separate types of reasoning on which human decisions tend to be based: deliberative and automatic.[144] The high speed at which RAS could perform analyses in order to keep up with technological developments in warfare increasingly requires humans to supervise or make decisions on whether to follow a system's 'judgment' using automatic reasoning rather than longer, more weighed deliberative reasoning. The surpassing of human cognitive abilities together with the knowledge that lives may be in danger from a mistake, can result in a sense of urgency that further affects human judgment on whether or not the correct conclusions have been drawn by the system. This may effectively leave certain originally human decisions up to machines, and ultimately it may warp what is considered an absolute necessity or an 'imminent

[141] Melzer, "Keeping the Balance between Military Necessity and Humanity – a Response to Four Critiques of the ICRC's Interpretive Guidance on the Notion of Direct Participation in Hostilities," 833; Schmitt and Thurnher, "'Out of the Loop': Autonomous Weapon Systems and the Law of Armed Conflict."

[142] See Scenario 4 'Back to basic' for the full description of the situation.

[143] "Reframing Autonomous Weapons Systems."

[144] Noel Sharkey does this based on human psychology research, which "divides human reasoning into two types: (i) fast *automatic* processes needed for routine and/or well tasks like riding a bicycle or playing tennis and (ii) slower *deliberative* processes needed for thoughtful reasoning such as making a diplomatic decision." See Sharkey, "Guidelines for the Human Control of Weapons Systems," 2.

threat'. This is especially relevant to decisions concerning whether or not to use kinetic force, whether defensively or offensively.[145]

Distinction

The third principle to be adhered to is the notion that there must be a distinction between legitimate (e.g., active military and combatant) and non-legitimate targets (e.g., civilians, civilian objects, surrendering soldiers, or medical staff). This distinction lies at the core of the regulation of hostilities.[146]

For the purposes of this paper's topic, the principle implies that there must be a sufficient level of certainty that RAS can ensure distinction between different 'types' of actors in a potential zone of action. In practice this means that militaries may only use weapons that can distinguish valid targets from civilian or protected targets. Presently, lacking autonomous weapons with such advanced abilities, militaries must have human decision-making in place at all points in the targeting process where it is necessary to ensure the principle of distinction is upheld.

On the topic of distinction there is much debate surrounding the abilities of RAS, as even for humans it can be challenging to distinguish between combatants participating directly in hostilities and the locations or buildings associated with them.[147] Existing AWS can only 'know' the difference between military targets and civilian objects under particular circumstances and in particular environments. However, training with systems that have learning abilities may be able to enhance their adaptivity and the possibility to prepare for more varied scenarios than initially programmed. Furthermore, similar to many other technologies or weapons, RAS are to be deployed for specific contexts and aims, as not all systems or programs fit all aims. At least for now though, RAS identify targets and warning signs based on certain, pre-programmed criteria, whereas conflict situations are unpredictable, and the identification of combatants does not usually adhere to easily programmable criteria. It is therefore unlikely that in the foreseeable future it will be possible to have a (weapon) system autonomously identify valid military targets across a variety of contexts.[148]

[145] Schwarz, "The (Im)Possibility of Meaningful Human Control for Lethal Autonomous Weapon Systems."

[146] Netherlands Advisory Council on International Affairs, "Autonomous Weapon Systems: The Need for Meaningful Human Control."

[147] For countries' description of combatants, see "Customary IHL - Practice Relating to Rule 14. Proportionality in Attack." For there to be direct participation in hostilities, there are three criteria: "a threshold of harm, a causal link between the act and the harm, and a connection to one of the parties to an armed conflict", Netherlands Advisory Council on International Affairs, "Autonomous Weapon Systems: The Need for Meaningful Human Control," 25.

[148] Netherlands Advisory Council on International Affairs, "Autonomous Weapon Systems: The Need for Meaningful Human Control," 24–25.

Humanity

The last of the four principles in IHL discussed in this paper, is the principle of humanity. While in and of itself being a principle that enhances—and often limits—the aforementioned three, there are a number of elements that make humanity a standalone principle.[149]

In an important International Court of Justice (ICJ) advisory opinion, the prohibition of unnecessary suffering is made explicit.[150] The notion that the "employment of arms which uselessly aggravate the sufferings of disabled men, or render their death inevitable" would be "contrary to the laws of humanity" has been a core tenet of international law governing hostilities dating back as far as 1868.[151] It is therefore crucial in the development of RAS to keep this overarching principle in mind.[152]

The main point of debate surrounding the principle of humanity that is relevant to this paper, is whether or not military RAS will violate the 'Martens Clause' under IHL.[153] This clause can be found in different forms throughout various IHL treaties, but is most often quoted as follows, from Article 1(2) of Additional Protocol I to the Geneva Conventions:

> In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience.[154]

There are different interpretations of this, more often than not depending on the status of the interpreting actor and their stakes in a conflict.[155] The limiting interpretation is that the Martens Clause can be a legal argument—of customary international law—for the prohibition of certain actions or, in the case of RAS, certain systems.[156] The opposing, permissive interpretation is that this clause is relatively insignificant, and simply functions as reaffirming signatories' acknowledgment of governance by customary international law. The third, more middle-ground reading of the clause is that it can function as a legal argument to present the illegality of

[149] Davison, "A Legal Perspective: Autonomous Weapon Systems under International Humanitarian Law."

[150] Legality of the Threat or Use of Nuclear Weapons (Advisory Opinion of 8 July 1996), 1996 Reports paragraph 78.

[151] Declaration Renouncing the Use, in Time of War, of Explosive Projectiles Under 400 Grammes Weight.

[152] Declaration Renouncing the Use, in Time of War, of Explosive Projectiles Under 400 Grammes Weight.

[153] See Hughes, "No, Autonomous Weapon Systems Are Not Unlawful under the Martens Clause"; Docherty, "Banning 'Killer Robots': The Legal Obligations of the Martens Clause"; Asaro, "Jus Nascendi, Robotic Weapons and the Martens Clause."

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

[155] Sparrow, "Ethics as a Source of Law."

[156] Docherty, "Losing Humanity."

certain systems, either alone or in conjunction with other legal arguments, but is itself not inherently a prohibition of any action or weapon per se.[157]

There has been debate over the extent to which the Martens Clause applies in the case of military RAS. On the one hand, the clause may apply to AWS because these are not addressed specifically by international law.[158] On the other hand, although neither RAS or AWS are specifically mentioned, their use and the limits to this use are dictated by IHL and the weapons review regulations of customary international law.[159] These different viewpoints will be expanded upon in Chapter 4.2, as they are part of key discussions on AWS developments' effects on human dignity. Most importantly though, the Martens Clause prevents "the assumption that anything not explicitly prohibited is permitted," and thereby has the ability to act as legal grounds for various policy directions taken in "new situations and new means and methods of warfare."[160] The notion from the Martens Clause that the development of new technologies such as RAS depends in part on the dictates of public conscience, makes societal debate an important consideration for the militaries of democratic nations in deciding whether or not, and how, to develop military RAS.

4.2 Dignity in the use of Autonomous Weapon Systems

Linked to the principles of IHL discussed above, as well as to the earlier topic of meaningful human control, a debate exists among academics as well as (inter)national policymakers concerning the basic understanding of what is most 'ethical' or 'humane'. While future capabilities of RAS might make certain missions easier for the military, or may end up helping save lives in conflict,[161] their use can still be considered as diminishing certain inherently human aspects of conflict. There is, however, no set definition of what the most dignified way is to go about preventing suffering, making it difficult to identify a widely accepted explanation of what is or is not 'humane' and 'dignified'. The two sides arguing this debate tend to operate on a different level of understanding of the notion of 'humane': on the one hand, it is said that decisions of life and death are inherently human and delegation thereof to a machine would be per definition inhumane; on the other hand, it is said if human suffering can be reduced, then not doing so would be inhumane. To illustrate this with a hypothetical: if using a certain type of RAS were to decrease the number of civilian deaths from twenty to ten in a particular conflict situation, one could either argue that the inherent fact that RAS use led to ten people dying is counter to human dignity, or one could argue that

[157] Cassese, "The Martens Clause: Half a Loaf or Simply Pie in the Sky?"

[158] Docherty, "Losing Humanity"; Docherty, "Heed the Call."

[159] Press, "Of Robots and Rules: Autonomous Weapon Systems in the Law of Armed Conflict."

[160] Davison, "A Legal Perspective :Autonomous Weapon Systems under International Humanitarian Law," 8.

[161] Arkin, "Lethal Autonomous Systems and the Plight of the Non-Combatant."

RAS made the situation more dignified because IHL was better observed, as ten less people died than otherwise would have been the case.

The following two sub-sections present the arguments for both sides of the debate on human dignity in warfare. The first sub-section presents the argument that the introduction of RAS will undermine dignity, while the second sub-section discusses how RAS may strengthen the position of human dignity in warfare. This discussion is presented as a debate in order to lay out the key arguments that have been presented to support or oppose development and deployment of military RAS in both the near and the more distant future.

Perspective One: Undermining Dignity

As the realm of machine learning expands, militaries stand to gain, both defensively and offensively, in terms of speed and efficiency. Automated defenses are not particularly new—think back to the earlier example of Israel's Harpy system or the US' Aegis weapon system—but future RAS may go further than protection and venture into the realm of counter-attacks.[162] Not only this, but machine learning algorithms will increasingly be employed to help inform decisions on resorting to the use of force internationally.[163] This of course raises issues about whether such systems provide a reliable analysis, and to what extent a human operator can question or overrule the recommendations produced through AI systems. As the value of such algorithms resides in their speed and the possibility to react immediately if it is deemed necessary for self-preservation, "the temptation to rely on the algorithm alone to guide decision-making [...] will be powerful."[164]

As will be discussed in the following paragraphs, primary arguments against RAS in the context of human dignity pertain to decisions on the use of force, distance, explainability, and threats to peacebuilding.

An issue brought up most often is that the barriers to using force may be reduced by the increased use of RAS, forming a problem for the general rule that the use of force should be a last resort, intended as self-defense.[165] This lowering of thresholds to violence could happen in different ways. For one, if there are fewer human lives endangered on the side of the attacker, the risk posed by the possible outcomes of an attack is lowered. Alternatively, it could happen that an early warning system is

[162] Deeks, Lubell, and Murray, "Machine Learning, Artificial Intelligence, and the Use of Force by States."

[163] Deeks, Lubell, and Murray.

[164] Deeks, Lubell, and Murray, 10.

[165] UN Charter Article 51; ICJ, Nicaragua Case (Merits), para 191: only the most grave forms of attack qualify; para 176: "self-defence would warrant only measures which are proportional to the armed attack and necessary to respond to it."

spoofed[166] by another (state) actor into 'recognizing' imminent danger when certain pre-programmed boxes are ticked, thereby setting off a counter-response.[167]

On the issue of distance, recent decades' development in military involvement abroad and the changing nature of this type of conflict form an interesting backdrop to the further integration of RAS in the military. In the case of the US, the international operations it has carried out over the past years, with less 'boots on the ground' and a larger role for military contractors, enabled the country to "maintain the appearance of a small military footprint with minimal risk of harm to US troops."[168] Throughout interventions, from Libya to the latest intervention against ISIS in Iraq, this has paved the way for the legal groundwork to claim that having fewer troop casualties makes military involvement a more valid option in the US' foreign policy.[169] As more functions of soldiers can be performed at a distance and/or by an autonomous system, the argument on a national level that military action abroad is more acceptable if the country's own troops are in less danger, is strengthened. The fear here is that this physical and moral distance from the battlefield or conflict zone will further lower the barrier to the international use of force. The lines drawn to regulate to the use of force are further blurred by the prevalence of involvement by proxy; asymmetric forces; the lack of clarity on which actors or parties are involved; whether hostilities constitute a full-blown war; and where the geographical 'border' of this war should be. If a situation cannot be classified as war, then per definition IHL cannot be upheld: IHL governs warfare.

Aside from the above, the distance between the operator and the target is another point of concern—not dissimilar to discussions seen during the deployment of drones or even the initial use of airpower.[170] There is an assertion that human dignity is undermined if machines effectively have the last say in who lives or dies—be it on purpose or by accident. With the increased use of automation and autonomization, two forms of distance have a possible impact on operations: institutional and physical.[171] As for the former, operating within an ethical armed force is not just about outcomes, but also about the processes that led there. Fully autonomous weapons "would lack the human judgment necessary" to ensure all agreements and obligations

[166] It appears that not only other actors hacking a country's system is a realistic risk, but so too is 'spoofing', or the tricking of system algorithms, often by mimicking patterns known to set off a certain reaction in the system. Extensive experiments with AI image-classification algorithms have shown that these systems are easily tricked with relatively small deviations from standardized representations. For example, one such experiment saw a turtle continuously being misidentified as a rifle. See Klare, "Autonomous Weapons Systems and the Laws of War."

[167] Deeks, Lubell, and Murray, "Machine Learning, Artificial Intelligence, and the Use of Force by States."

[168] Dickinson, "Drones, Automated Weapons, and Private Military Contractors," 111.

[169] Dickinson, "Drones, Automated Weapons, and Private Military Contractors."

[170] Banta, "'The Sort of War They Deserve'?"

[171] Feickert et al., "U.S. Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress."

are followed through in a way that can be justified after the fact.[172] And as for the latter, even with a human closely monitoring or making certain decisions, the physical distance created by (semi-)autonomous systems can also lead to a "moral distance as the face of the opponent becomes less visible, which eliminates the moral–psychological barrier for killing."[173]

The explainability of the decision-making process—as described in Chapter 3.1—is closely linked to the previous point on the distance of the human from the battlefield. Using RAS not just on an operational or tactical level, but incorporating such complex systems into wider decision-making processes will add to the already prevalent transparency issues that the military has toward the public—and transparency is a "key element in enabling society to have the right amount of trust and confidence in the operations of an AI system."[174] Along with a lack of transparency into decision-making comes less scrutiny of the quality and interpretability of machine-produced recommendations or predictions upon which decisions are based, and therefore it may also lead to a lessened sense of responsibility.[175] Aside from the possibility that for this reason a state may be less willing to explain machine reasoning behind decisions, another is that a state may unable to explain it—a possibility with far more consequences for the position that human dignity considerations hold in policy-making.[176]

The final point of discussion is a set of arguments pertaining to the post-war effects of military RAS on human dignity. While increasingly the world sees conflicts of various intensity with no clear start or end, the feasibility of peacemaking and peacebuilding should nevertheless always be on the mind of involved states. Visible use of RAS may undermine counterinsurgency efforts intended to stabilize a region, in a way similar to how civilians can react to or fear the use of force by armed drones. A military's intention may be good, but all that is heard or seen on the ground is the noise of an overhead drone and the resulting destruction.[177] At this point in time, for similar reasons, RAS "won't help win the hearts and minds of the occupied or vanquished," and may make it more difficult to achieve lasting peace.[178] RAS use can be perceived on the ground as a lack of commitment if used by a state making peacebuilding efforts. Aside from this, it may also hamper partnership efforts. This is a point

[172] Davison, "Autonomous Weapon Systems"; "Killer Robots - Learn."

[173] Verdiesen, "Agency Perception and Moral Values Related to Autonomous Weapons," 14.

[174] Charisi et al., "Towards Moral Autonomous Systems."

[175] Johansson, "Ethical Aspects of Military Maritime and Aerial Autonomous Systems."

[176] Feickert et al., "U.S. Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress."

[177] Khan and Gopal, "The Uncounted".

[178] Lin, Bekey, and Abney, "Autonomous Military Robotics: Risk, Ethics, and Design."

especially relevant to forces—like the RNLA[179]—which have an approach to peacebuilding that takes into account the importance of understanding the area of operations and the local sensitivities that can make or break the success of such missions. RAS are incapable of developing or replacing the personal relationships with the local population that are generally necessary for trust and, by extension, for a successful mission.[180]

Perspective Two: Aiding Dignity

Several points run counter to the above discussed argument that military RAS would lower barriers to the use of force.[181] The first is the fact that as the speed of (cyber) attacks steadily increases, so does the speed with which a response must be readied. Machines could calculate endless different options, outcomes, and consequences at far greater speed and accuracy than humans, and could therefore improve both decisions on self-defense and counter-attacks, as well as on legitimate targets or intensity of attacks. This is still dependent on the possibility of satisfying the requirements of proportionality, necessity, and distinction. Another argument is that the public knowledge of the fact that certain states have certain capabilities may work as a deterrent.[182] This was also noted in scenario 4 of the expert session, where the autonomous system and its parameters that were communicated to the adversary's higher command limited the strategic choices of the adversary, as a defensive AWS is more likely to be consistent in its behavior than a human operator.

Another common argument is that distance will also lower the threshold to using force. However, through the use of a semi-autonomous system operated at a distance, reduced stress levels and increased evaluation time may allow a human operator the opportunity to make better informed and as a result, more ethical decisions.[183] It can be argued that if a reliable and IHL-observant autonomous system is developed, for any relevant type of operation, it should perhaps "not only be regarded as morally acceptable but also [...] ethically preferable over human fighters" if unethical situations can thereby be avoided.[184]

An additional reason that RAS could in the future add to human dignity, is the lack of human shortcomings, such as fatigue or emotions clouding judgment. It is possible to

[179] van der Lijn and Ros, "Peacekeeping Contributor Profile: The Netherlands."

[180] Marchant et al., "International Governance of Autonomous Military Robots," 2889.

[181] Feickert et al., "U.S. Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress."

[182] Deeks, Lubell, and Murray, "Machine Learning, Artificial Intelligence, and the Use of Force by States."

[183] Davison, "Autonomous Weapon Systems"; Arkin, *Governing Lethal Behavior in Autonomous Robots*, 47–48; Strawser, "Moral Predators"; Leveringhaus, "Autonomous Weapons Mini-Series."

[184] Etzioni and Etzioni, "Pros and Cons of Autonomous Weapons Systems," 74.

embed a mission's rules of engagement (ROEs) in a system, and once technological development is far enough that RAS can observe IHL, this could prevent mistakes or even serious harm that may otherwise arise under the strenuous circumstances soldiers are often in.

With respect to the discussion on the difficulty of explaining the actions of RAS, results from machine-produced recommendations may make it easier for states to lay out the rationale that led to a certain outcome—rather than a decision being a commander's intuition.[185] A difficulty here is that much of what goes on in algorithmic functions is a 'black box', and while some systems log decision-making and errors, with increasingly complex systems it can be difficult to establish precisely how some results were reached. This topic was discussed in depth in Chapter 3.1.

Overall, there is still a long way to go before machines could autonomously satisfy IHL requirements. However, certain examples can show that there are rapid developments in this field, sped up by the blurring lines between civil and military R&D. Such overlap may prove necessary to keep up with the way in which war is changing. The world is becoming increasingly urbanized, with over two thirds of the world projected to live in cities by 2050,[186] creating complex security problems for local governance that can escalate and become local or regional conflicts.[187]

At the same time, the trend of conflicts taking place in cities is only set to continue,[188] and the packed, hard to navigate, and 'easy-to-hide-in' nature of cities, means that the principles of IHL will only become harder to adhere to for soldiers in high-intensity situations.[189] This means that states developing RAS have a great interest in further research into RAS capabilities that will improve their militaries' ability to withstand the IHL test in complicated environments.[190] The US, for example, uses war games in fictitious cities to run through scenarios and establish what types of technology it will need to get through the coming decades, knowing that all war-fighting functions "are complicated and challenged by the compartmentalized terrain that's present in the urban environment."[191] RAS may provide a crucial tool enabling militaries to have the

[185] Deeks, Lubell, and Murray, "Machine Learning, Artificial Intelligence, and the Use of Force by States."

[186] United Nations, Department of Economic and Social Affairs, Population Division, "World Urbanization Prospects: The 2018 Revision."

[187] Horowitz, "Joint Blog Series: Precautionary Measures in Urban Warfare: A Commander's Obligation to Obtain Information."

[188] "Preparing for More Urban Warfare."

[189] Horowitz, "Joint Blog Series: Precautionary Measures in Urban Warfare: A Commander's Obligation to Obtain Information."

[190] A way RAS can improve militaries' functioning in such difficult environments is, for example, a tool like Hivemapper. This "creates 3D maps from videos captured by drones, aircraft, and ground vehicles" as a way of "using machine learning tech to augment human analysts." See Weisgerber, "What's in the House NDAA?; Pentagon's 3D-Mapping Service; New Marine One, Weed Whacker; and More."

[191] Apart from urban environments created to train civil protection units with autonomous systems, the US Defense Department also uses a specially built city for military training purposes. The idea is to lead the way in innovation of

intelligence and preparation needed to better assure all precautions are taken and all IHL principles are upheld in the complicated modern war scenarios.

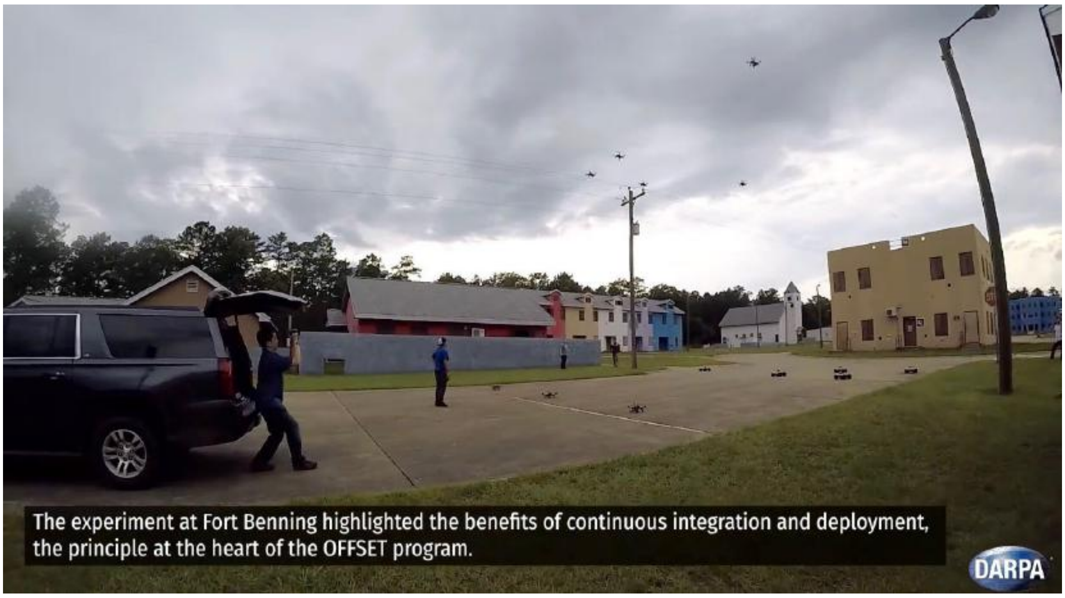
The experiment at Fort Benning highlighted the benefits of continuous integration and deployment, the principle at the heart of the OFFSET program.
DARPA

Figure 9. Swarm Autonomy Test for DARPA's OFFensive Swarm-Enabled Tactics (OFFSET) program at Fort Benning, Georgia, August 2019. Photo: US DARPA[192]

4.3 Summary

Crucial to upholding human dignity in conflict is adherence to and respect for International Humanitarian Law. The most important principles that RAS should be able to respect are proportionality, military necessity, distinction, and humanity (often known as the prevention of unnecessary suffering). Recent developments of both RAS in general and AWS in particular are making strides in this regard. They can 'distinguish' to a certain extent the difference between military targets and civilian objects, albeit only under particular circumstances. Training with systems that have learning abilities may be able to enhance their adaptivity and the possibility to prepare for far more different scenarios than initially programmed. RAS can greatly reduce the amount of time needed to come to conclusions or to make decisions. At the same time, however, surpassing human cognitive abilities strongly affects the mitigating role of human judgment in assessing whether or not the correct conclusions have been drawn by the system. Even so, RAS capabilities can also improve militaries' ability to navigate the complicated environments in which modern conflicts tend to be fought, by improving intelligence, logistics, evacuation and other capabilities

swarm system capabilities to “assert and maintain superiority of the urban operating environment.” See Chung, “OFFensive Swarm-Enabled Tactics (OFFSET)”; Peters, “Watch DARPA test out a swarm of drones”; Calloway, “Army Wargames Shape the Future of Urban Warfare”; Musgrave, “Inside ‘Liberty City,’ Homeland Security’s Site for Testing Urban Drones.”

[192] DARPAtv, “Teams Test Swarm Autonomy in Second Major OFFSET Field Experiment”.

crucial to effective mission functioning. For the foreseeable future, however, it will not be possible to have an offensive (weapon) system that can autonomously distinguish targets to the standards of IHL and weigh all the relevant contextual information as well as a human.

The debate on the effects of integrating RAS into the armed forces is one with a wide range of arguments. The ongoing academic debate is important to ensure policymakers can make informed decisions that will affect the future of new technologies like RAS. Meanwhile, Machine Learning and Artificial Intelligence are becoming more common in shaping decisions in military contexts. It is also likely that intelligent system design will continue to become more intricately woven into the fabric of decision-making—not only on the battlefield, but also in the policy-making world that governs it. In the future, this development will continue to make the attribution of wrongful use of force and the assigning of responsibility more complicated. At the same time, RAS may provide a crucial tool in improving militaries' intelligence and preparations that ensure IHL principles are upheld in increasingly complicated modern warfighting scenarios.

5 | Responsibility and Accountability

The final component of ethics in the context of this study concerns the establishment of accountability and responsibility before, during, and after the deployment of RAS in a military context. This chapter will re-examine the three perspectives of meaningful human control (MHC) introduced in Chapter 3, namely the life cycle, sub-system and OODA loop perspectives (See Figure 4). This approach highlights all the elements relevant to maintaining MHC and is useful in illustrating the numerous actors involved in the life cycle of RAS and all the points at which they may be (partially) responsible for certain courses of action. The approach also exemplifies the importance of considering accountability beforehand, in order to ensure it is clear who shares responsibility throughout the long and relatively fragmented chain of R&D and usage that typify military technology.

The discussion on military RAS often seeks to compare humans and machines. An example of this is "the difference between a pilot flying an airplane on autopilot and an airplane with no human in the cockpit at all."[193] This chapter on accountability and responsibility does not dwell on whether the former or the latter way of flying is 'better'. The question is, rather, in the case of system or human failure, what amount of damage caused by failure will be deemed acceptable by the military or society? How can this be assessed early on and how can the risks associated with flying on autopilot or autonomously be mitigated, in line with that baseline of ethical standards?

Currently, there are obstacles to determining responsibility and establishing accountability for activities involving RAS.[194] First, Chapter 5.1 goes into the current ways in which accountability may apply in the case of RAS from a legal point of view,[195] especially in cases where wrongdoing occurred unintentionally. The legal requirement of intent behind action presents a difficulty, as it is not always clear whose intention should be reckoned with when it comes to the deployment of RAS, nor how this intent could be sufficiently established in a legal sense in the first place. Chapter 5.2 discusses whether existing legal and accountability frameworks are sufficient to safeguard society's ethical standards, and it discusses the existence of a legal accountability gap, both now and in the future. Chapter 5.3 explores improvements to or developments in addressing the challenges of accountability for RAS, and addresses how to understand responsibility and accountability in a non-legal, institutional way.

[193] Scharre, *Army of None: Autonomous Weapons and the Future of War*, 193.

[194] Schmitt and Thurnher, "'Out of the Loop': Autonomous Weapon Systems and the Law of Armed Conflict."

[195] In legal terms, 'responsibility' refers to a duty to act with due diligence. 'Accountability' refers to "the process aimed at a [...] public assessment of [...] conduct in a given case in order to evaluate whether this conduct was required and/or justified" based on established responsibility. Finally, the term 'liability' follows this, and refers to the attachment of legal consequences to said conduct.

5.1 Current Frameworks of Legal Accountability for Military Practices

Integrating RAS into military operations may erode moral responsibility, as repercussions for IHL non-compliance requires legal individual and/or collective accountability. If one would seek some form of accountability for certain outcomes of RAS' actions or decisions, there must be shown to be a link between the outcome of a RAS-dictated action and the intent of those responsible for its development and/or operation.[196] First, the discussion on legal personhood of RAS is set out by briefly laying out recent years' developments in this field, in particular developments in Europe.[197] After this, the chapter discusses the two primary means through which legal accountability and liability can be established, namely state and criminal responsibility, as well as the shortcomings tied to these legal frameworks in the governance of RAS. Lastly, the chapter introduces a number of alternative bodies of law that have been proposed to incorporate or take from in establishing a legal framework to deal with possible RAS uses in the future.

Legal Personhood and Autonomous Systems

With RAS at its present stage of development, in the ethical and legal sense, responsibility and accountability for their actions fall upon humans.[198] Although legal personhood already exists—for international organizations and companies, for example—, the European Parliament has proposed to the European Commission to consider extending this to a form of legal, "electronic" personhood for robots.[199] This suggestion has received mixed reviews.[200]

There are several points concerning the notion of legal personhood for autonomous synthetic entities. First and foremost,

> [t]he basic provisions for a legal person are: 1. that it is able to know and execute its rights as a legal agent, and 2. that it is subject to legal sanctions ordinarily applied to humans.[201]

[196] International Committee of the Red Cross ICRC, "Ethics and Autonomous Weapon Systems"; Feickert et al., "U.S. Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress."

[197] See Committee on Legal Affairs, "Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))."

[198] Shilo, "Speaking of Responsibility: Autonomous Weapon Systems, State and Individual Responsibility"; Lin, Bekey, and Abney, "Autonomous Military Robotics: Risk, Ethics, and Design."

[199] The suggestions made by the Parliament include a definition of what characteristics would constitute 'smart' autonomous robots. See Committee on Legal Affairs, "Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))."

[200] See e.g., Bryson, Diamantis, and Grant, "Of, For, and By the People: The Legal Lacuna of Synthetic Persons"; Vincent, "Giving Robots 'Personhood' Is Actually about Making Corporations Accountable."

[201] Bryson, Diamantis, and Grant, "Of, For, and By the People: The Legal Lacuna of Synthetic Persons."

Second, 'legal personhood' is a technical term that does not necessarily imply somehow viewing robots as inherently human or ethical actors. Rather, the term gives way to any number of rights and obligations, because it means that a legal system addresses its rules to the actor or entity.[202] The extension of the term 'legal personhood' is largely a legal tool of convenience within civil law, in the same way legal personhood is given to a company in order to provide them with legal rights as well as obligations. While much discussion on the European Parliament's report focused on the notion of 'electronic personhood', the aim of the report's recommendations was more to ensure that establishing "a causal link between the harmful behavior of the robot and the damage suffered by the injured party" could become sufficient to claim compensation from a company.

The autonomous robots envisioned in this Parliamentary text were for civil use rather than military. However, anticipating future risks posed by increased autonomy and addressing the liability gap that may arise if no legislative care is taken to address this topic, is equally relevant for military RAS. More on civil law instruments that could be relevant for military RAS is covered further on in this chapter.

State Responsibility

Responsibility for state wrongdoing is established on the basis of Article 2 of the International Law Commission's (ILC) Draft Articles on the Responsibility of States for Internationally Wrongful Acts, which dictates that

> there is an internationally wrongful act of a State when conduct consisting of an action or omission: (a) is attributable to the State under international law; and (b) constitutes a breach of an international obligation of the State.[203]

However, the state itself as an entity is made up of the people that represent it, meaning "an 'act of the State' must involve some action or omission by a human being or group [...]."[204] Furthermore, the "only conduct attributed to the State at the international level is that of its organs of government, or of others who have acted under the direction, instigation or control of those organs, i.e., as agents of the State."[205] At this point a distinction can be made between on the one hand agents of the state that take orders within an explicitly established command structure, and on the other hand agents of the state that make an individual decision that results in wrongful act, outside of an 'effective command and control' structure.[206] The latter

[202] Bryson, Diamantis, and Grant.

[203] International Law Commission, "Responsibility of States for Internationally Wrongful Acts."

[204] International Law Commission, 35, paragraph 5.

[205] International Law Commission, 38, paragraph 2.

[206] "Killer Robots and the Concept of Meaningful Human Control."

will be discussed further in the criminal responsibility section. In summary, a breach of international law has a human link and requires the presence of humans, which at the state level manifests itself through agents of the state.

States have the duty to respect and ensure compliance with IHL under Common Article 1 of the Geneva Conventions.[207] Moreover, state obligations include the regulation of companies to ensure that emerging technologies are not in violation of IHL.[208] This obligation is extended under Article 36 of the Protocol Additional I to the Geneva Convention, whereby

> [i]n the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.[209]

The state is therefore made responsible for testing and certifying RAS provided by domestic private contractors and/or foreign suppliers. Compliance with Article 36 requires integration of international obligations throughout most of the life cycle of RAS, from design and procurement to its adoption by the military. Developing and institutionalizing a system such as the example presented by Australia in Chapter 5.3 could be a way of ensuring that there is state responsibility where it is required under a nation's international obligations. Without a clearer overview of where and to what extent it can be reasonably expected for a state's responsibility to lie the research and development of RAS, it is difficult to hold states accountable for the consequences of their use.

Criminal Responsibility

A cornerstone of international criminal law (ICL) is the attribution of individual criminal responsibility. For this body of law to be applicable, there must be criminal intent (*mens rea*) involved, or, as is generally the case for war crimes, the wrongful act must have been committed "willfully."[210] This criminal intent is what separates civil and criminal law. Someone who drives a car into a pedestrian with the intention to

[207] "Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field. Geneva, 12 August 1949."

[208] This type of obligation is known as a 'positive obligation', meaning states must make an active effort to ensure compliance with the law, such as adopting (new) measures to uphold it. A 'negative obligation' means simply to refrain from certain acts that would violate the law in question. These terms are most used in International Human Rights Law.

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

[210] Crootof, "War Torts: Accountability for Autonomous Weapons." "Willfully" here means that someone must have acted either intentionally or recklessly, see Sandoz, Swinarski, and Zimmermann, "Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949."

hurt or kill will be liable under criminal law. However, someone who loses control of their car, cannot brake in time, and hurts a pedestrian in the ensuing accident, will be liable under civil law and will most likely end up paying monetary damages as a result.

ICL cannot be applied to RAS directly. As RAS have no consciousness, no criminal intent can be established. One could also question whether this would even make sense in the first place, as the purpose of ICL is to establish willful wrongdoing and appropriate punishment, and there is not much effectiveness in applying human punishments to machines. Hence, for there to be criminal accountability under certain circumstances, an individual, or group of individuals, will need to be held responsible. Assigning individual responsibility under ICL for the use of RAS/AWS will be difficult, however. This is largely due to the still inadequately agreed upon concept of meaningful human control, as well as the increase in decision-making by or based on machines coupled with the degree of unpredictability that still exists at the current stage of technology's development.

Criminal responsibility can—in theory—be attributed to individuals in all phases of the RAS life cycle, from programmers to operators. In practice, it would prove complicated to pinpoint one or several culprits among the many people involved throughout RAS life cycles. A RAS criminal case would be relatively 'easy' to solve if it were possible to trace back the machine-produced wrongful act to certain people, for example, if there was deliberately incorrect pre-configuration by a programmer, or recklessness on the part of an operator. However, there are two key difficulties here. First, there is the sheer number of people tasked with building an algorithm, manufacturing and assembling RAS hardware, developing training and evaluation, and all other steps involved in RAS research and development. This argument is visible in the Boeing 737 MAX 8 example discussed in Chapter 3.1. Second, most issues in applying ICL to RAS operations arise from oversight or mistakes, rather than intentionally wrongful acts, meaning no one can be held directly liable. The fact that RAS may result in serious violations of IHL without human intention as the main driving or facilitating factor, seriously hampers attaining justice in the case of wrongful acts. The diffusion of responsibility means it becomes more likely that no one will be punished even in light of a mistake with lethal consequences as a result of the use of RAS.

There are various doctrines within ICL, which cannot all be discussed here, but some of which may be relevant in assigning responsibility in a military context. Not all ICL doctrines require every participant in a crime to have intended this crime, but they are still all "premised on the notion that there is at least one individual who did possess the requisite intent."[211] One of these doctrines often cited as being most relevant in when it comes to military RAS is that of command responsibility. In this doctrine, a commander can be held legally responsible for the actions of a subordinate if they had "effective command and control, or effective authority and control over the forces that committed the crime."[212] In a ruling by the International Criminal Tribunal for Rwanda, it was established that the "material ability to control the actions of subordinates is the touchstone of individual [command] responsibility."[213] In other words, this responsibility is conditional upon not only the subordinate(s) intent to commit a wrongful act, but also upon the commander's material ability—or lack thereof—to actually prevent and punish the commission of the offense.[214] If it is proven that a commander was not realistically in a position to prevent or punish the actions of RAS/AWS, this means it is unlikely that the commander would be criminally liable. This is the case since criminal responsibility is attributed after the fact (*ex post*), while the use of AWS is permitted before the event (*ex ante*) under the CCW and in compliance with IHL.[215] Moreover, in the case of supervised autonomous systems with the possibility of human override, rather than see their well-intentioned intervention resulting in a negative outcome, an operator may instead prefer to benefit from plausible deniability after inaction. Thus, there is an adverse incentive for the operator overseeing the RAS not to intervene if mistakes may be met with criminal liability afterward.[216]

The issue with attempting to assign criminal responsibility, within any of the doctrines considered, is that it is inherently tied to the individual and their intentions. There are increasingly lengthy research and development processes for RAS, as well as more diffusion of tasks in these processes among government, military, and private sector actors, meaning that there are almost never specific individuals solely responsible for the consequences of RAS deployment. Moreover, where exactly state responsibility starts and ends is currently unclear for parts of the RAS life cycle, and individual criminal responsibility is often not applicable. Altogether, the existing legal frameworks surrounding responsibility for, and dealing with, mistakes as well as wrongful actions as a result of the use of military RAS appear to have limitations.

[211] Dickinson, "Drones, Automated Weapons, and Private Military Contractors," 116.

[212] "Killer Robots and the Concept of Meaningful Human Control"; Galand, Hunter, and Utmelidze, "International Criminal Law Guidelines: Command Responsibility," 65.

[213] Prosecutor v Kavishema, paragraph 229.

[214] Shilo, "Speaking of Responsibility: Autonomous Weapon Systems, State and Individual Responsibility"; Mucic et al., "Celebici", paragraph 378.

[215] Shilo, "Speaking of Responsibility: Autonomous Weapon Systems, State and Individual Responsibility"; Bo and Woodcock, "Blog: Lethal Autonomous Weapons, War Crimes, and the Convention on Conventional Weapons."

[216] Chehtman, "New Technologies Symposium."

Alternative Bodies of Law

Several bodies of law have been brought forward in attempts establish responsibility as well as make remedies for wrongful acts by AWS possible. The latter is relevant for cases where no human criminal intent can be established, but there was still wrongdoing. One such body of law is contract law, which has also been considered for regulating private military and security contractors (PMSC) more generally.[217] The Montreux Document, for example, suggests that States "include contractual clauses and performance requirements that ensure respect for relevant national law, international humanitarian law and human rights law by the contracted PMSCs."[218] Something could be said for contract law's possibility to circumvent jurisdictional obstacles to regulating wrongful acts in militaries dependent on private contractors, including for their RAS. Contract law may be a way to force private contractors to adhere to the norms of public international law.[219] There are issues with using private law to remedy public injustices, however. For one, regulating war crimes through the lens of contract law creates denial: "harm is cognitively reframed and then allocated to a different, less pejorative class of event"[220] until "a human rights violation is the same as a breach of contract."[221] Another issue is that all responsibilities or obligations are limited to precisely what is included in a contract's terms. Diving into the minutiae of one's precise obligations and the exact terms of a contract is quite normal in private law, but this inward-looking nature conflicts with the necessity in IHL to look beyond what is on paper.[222]

A second option that has been suggested introduces tort law.[223] Within common law, tort law is relevant for cases where there has been wrongdoing, but with no criminal intent behind the action. Whereas criminal law is set on prohibiting certain behavior, it has been suggested that tort law could offer "a means of regulating valuable but inherently dangerous activities and compensating injurious wrongs."[224] Where ICL is meant to hold individuals accountable for war crimes, "war torts" may form an added regime that could hold states accountable, circumventing the need for criminal intent. States are, in the end, responsible for making the choices that lead to the integration

[217] Liu, "Contract Law as Cover."

[218] "The Montreux Document - On Pertinent International Legal Obligations and Good Practices for States Related to Operations of Private Military and Security Companies during Armed Conflict", para A.IV.14 and 15 of Part II.

[219] Dickinson, "Contract as a Tool for Regulating Private Military Companies"; Dickinson, *Outsourcing War and Peace*, 69–101.

[220] Cohen, *States of Denial: Knowing about Atrocities and Suffering*, 106.

[221] Liu, "Contract Law as Cover," 24.

[222] Liu, 3.

[223] A 'tort' is a civil wrong causing loss or harm, which results in legal liability for the person who committed the tortious act. Tortious acts can range from inflicting emotional distress or financial losses, to inflicting injury or invading privacy.

[224] Crootof, "War Torts: Accountability for Autonomous Weapons," 1353.

of RAS into the military. The idea to primarily implement a tort regime with states in mind as the responsible actors therefore has come to the foreground the most.[225] Adding on to this notion of state responsibility in tort cases is the fact that "as long as a certain type of weapon is considered lawful and its production is ordered by a legitimate entity, corporate responsibility does not pose any contentious issues."[226] This is because manufacturers are absolved of liability if the system provided meets the legal conditions of the acquiring agency at the time of order.[227] A major issue with tort law is scalability.[228] While tort is a standard procedure in domestic legal systems, there is no international regime for it. Getting a regime off the ground that would allow people or groups to essentially sue states for damages incurred due to unpredictable RAS is difficult to envision.

5.2 A Legal Accountability Gap?

One of the key challenges in the use of RAS, and AWS in particular, is that the absence of full applicability of existing legal accountability frameworks, alongside inadequate agreement over what constitutes human control over a system, creates an 'accountability gap'. This means that in cases of violation of IHL, whether accidental or intended, there may be no human or entity directly responsible. Lacking clear establishment of responsibility, there may be no accountability for the actions of the RAS.[229]

To a certain extent, human responsibility for decisions on the use of weapon systems must be retained. Accountability and liability cannot be transferred to machines themselves, and this fact should push for consideration of who holds responsibility at many different points throughout the life cycle of RAS. The fundamental problem is the existing gap in international law is based on the permissible use of AWS under the CCW and the criminal responsibility attributed after the unlawful act involving AWS has taken place.[230] The formed discrepancy enables the operator to cite technical issues as the cause of the incident, leaving no one accountable for the actions of the AWS. Both the operators and to some degree the programmers are further distanced from responsibility through ambiguities resulting from third-party involvement in RAS development, software updating and self-learning abilities.[231] The issue was

[225] Malik, "Autonomous Weapon Systems: The Possibility and Probability of Accountability"; Crootof, "War Torts: Accountability for Autonomous Weapons."

[226] Malik, "Autonomous Weapon Systems: The Possibility and Probability of Accountability," 628–29.

[227] Boyle v United Techs. Corp. 487 U.S. 500, 510 (1988); Koohi v. United States, 976 F.2d 1328, 1336–37 (9th Cir. 1992)..

[228] Asaro, "The Liability Problem for Autonomous Artificial Agents."

[229] Horowitz and Scharre, "Meaningful Human Control in Weapon Systems," 8.

[230] Bo and Woodcock, "Blog: Lethal Autonomous Weapons, War Crimes, and the Convention on Conventional Weapons."

[231] Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," 181–82.

highlighted in the latest meeting at the CCW, where it was noted that "in the case of an incident involving LAWS, it was uncertain as to who would be held accountable within the chain of command or responsibility, such as the commander, programmer, or operator."[232]

What makes assigning responsibility for the actions of machines all the more challenging is the combination of RAS deployment and increased privatization of military materiel. Together, these developments have "fragmented decision-making over the use of force, rendering accountability for violations of IHL principles much more difficult to achieve."[233] Since the Nuremberg trials after the end of the Second World War, IHL has been a part of a decades-long trend toward individual responsibility, and it is a crucial aspect of IHL that perpetrators are held personally responsible if they commit wrongful acts. Yet autonomous weaponry and private contractors tend not to be situated in a military command structure, bringing decision-making and the consequences for its results "outside the ordinary bureaucratic chain of command."[234] The main problem posed by this, is that it becomes far more difficult to prove that a commander has the *de facto* level of control needed to demonstrate command responsibility. While the doctrine of command responsibility is possibly a better way to assign responsibility than attempting to find individual criminal intent behind a contracted programmer, it still has no concrete way of solving cases where wrongdoing has occurred as a result of multiple actors' actions, without these actors having intended this wrong.[235]

At the same time, the legal question that arose after WWII persists: the lack of clarity on the extent to which individuals can truly know what bigger picture their work is contributing to.[236] Therefore, the fact that firstly, "something more than ordinary negligence"[237] is the cornerstone of criminal responsibility; secondly, there remain a number of challenges to the establishing of state responsibility; and thirdly, other bodies of law have not yet been considered sufficiently in the context of RAS, means together that in light of increased privatization this body of law may prove to have significant limitations, especially when it comes to RAS.

[232] Chairperson of the Informal Meeting of Experts, "Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)," para. 52.

[233] Dickinson, "Drones, Automated Weapons, and Private Military Contractors," 96.

[234] Dickinson, 115.

[235] Dickinson, 118.

[236] D Luban, "Legal Modernism: Law, Meaning, and Violence," 372.

[237] A Danner, J Martinez, "Guilty Associations: Joint Criminal Enterprise, Command Responsibility, and the Development of International Criminal Law" 79.

5.3 Addressing the Gap

As has become clear, a central question concerns where responsibility lies or should lie in the operation of RAS, in general, but most importantly in cases of unintended and/or unlawful harm.[238] Addressing this question should include the consideration of the wide spectrum of actors involved in the development and deployment of military RAS, rather than just the operators in or on the OODA loop.[239] The increasingly dominant role of the private sector in the development of RAS has propelled this conversation forward. This highlights the difficulty of both establishing responsibility and accountability throughout the full R&D and deployment chain of RAS, as well as bridging the distance from the design of a single element within a system to the final (autonomous) 'decision'.[240]

The traditional, legal way of looking at responsibility, accountability, and liability is the following. In legal terms, 'responsibility' refers to a duty to act with due diligence. 'Accountability' refers to "the process aimed at a [...] public assessment of [...] conduct in a given case in order to evaluate whether this conduct was required and/or justified" based on established responsibility.[241] Finally, the term 'liability' follows this, and refers to the attachment of legal consequences to said conduct.

There is, however, another way of looking at these three terms that could be considered within the institutional set-up of RAS integration in the military. More so outside the realm of lawyers, it has become relatively standard in the public sphere to speak of 'liability' as being explicitly rules-based and 'responsibility' as governance-based. 'Accountability' should be something that comes before both as well as after, and it refers people's ability to explain and justify their behavior at all times.[242] This is a way of defining accountability that has also become quite standard practice in business spheres, where it is described as "[t]he obligation of an individual or organization to account for its activities, accept responsibility for them, and to disclose the results in a transparent manner."[243] It can be very difficult to establish direct responsibility due to the hierarchies of public governance, making it all the more important that accountability is emphasized in fields where third-party private companies and actors have such a crucial role, as is the case for military technologies such as RAS.

[238] Lin, Bekey, and Abney, "Robots in War: Issues of Risk and Ethics."

[239] Verdiesen, "Agency Perception and Moral Values Related to Autonomous Weapons"; Marra and Mcneil, "Understanding 'The Loop': Regulating the Next Generation of War Machines."

[240] Worcester, "Autonomous Warfare – A Revolution in Military Affairs."

[241] Giesen and Kristen, "Liability, Responsibility and Accountability: Crossing Borders," 6; Kool, "(Crime) Victims' Compensation: The Emergence of Convergence."

[242] See Folkman, "The '8 Great' Accountability Skills For Business Success"; Hoek, van Montfort, and Vermeer, "Enhancing Public Accountability in the Netherlands."

[243] "Accountability."

An example of how to institutionalize accountability and responsibility in this way was presented at the Governmental Group of Experts (GGE) meeting on the CCW in March 2019. Australia's representatives put forward a document describing how the country embeds 'control' into its weapon system development.[244] Australia's system of control "incrementally builds upon itself, embedding controls into military processes and capability at all stages of their design, development, training and usage," and at all stages reviews compliance with national and international legal obligations.[245] The stages of institutional development as presented by Australia, in their most simplified form, are as follows in Figure 10:

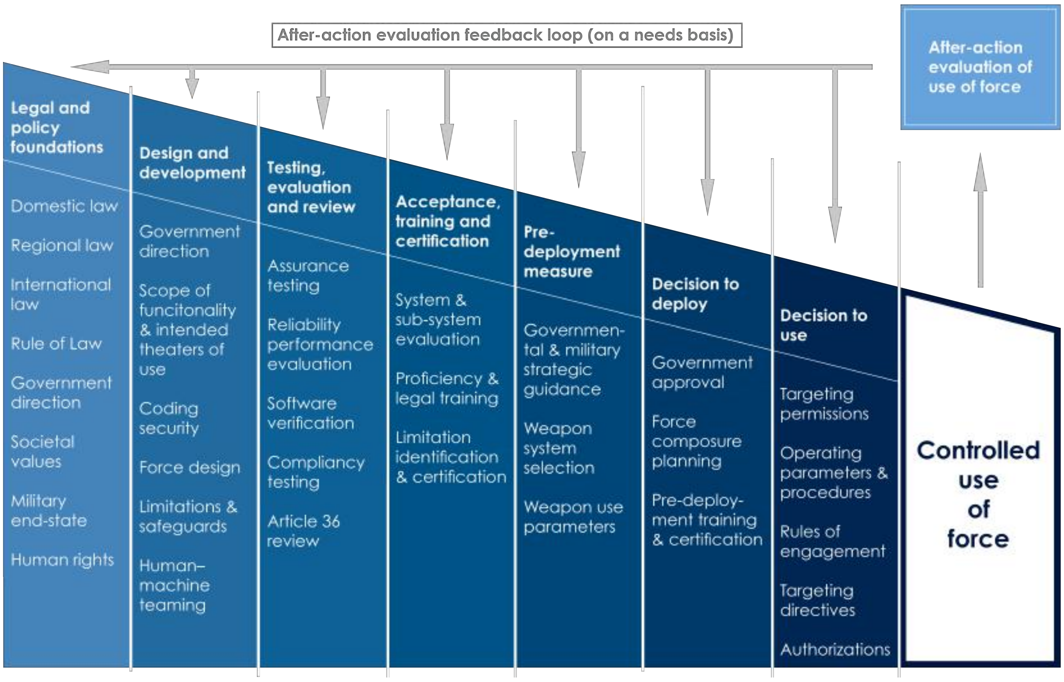

Figure 10. Australia's system of control for autonomous weapon systems[246]

Australia also used of this model of a 'system of control' to illustrate its view that the way the phrase 'human control' is often used does not always do justice to, or reckon with, practical military reality. Australia's representatives went so far as to argue that because the phrase 'human control' doesn't adequately cover military reality, it "does not provide a useful basis to further GGE discussions unless there were a common understanding of the term"[247] such as the delineation presented in Australia's model control system in Figure 10.

[244] In the document in question, the term 'control' referred to "the system of processes and procedures through which a state achieves its intended military effect, in a manner compliant with its legal obligations and policy objectives." See "Australia's System of Control and Applications for Autonomous Weapon Systems."

[245] "Australia's System of Control and Applications for Autonomous Weapon Systems."

[246] Adapted from "Australia's System of Control and Applications for Autonomous Weapon Systems."

[247] "Australia's System of Control and Applications for Autonomous Weapon Systems," 5.

Despite the fact that the details of this system of control may be up for debate, and although Australia's conclusion that such institutionalized control would solve AWS' responsibility issues is premature, this description of a state's development of AWS is very useful to ascertain key points in the processes of accountability. Such an approach can incorporate institutional assessment and constant evaluation of all three intrinsically linked concepts—life cycle, sub-system and OODA loop—which were introduced in Chapter 3 as being crucial to establishing sufficient control to act ethically when deploying and using RAS.

An important question, however, remains how far various states will be willing and able to realize such levels of control over private equipment manufacturers and contractors as the system visualized above requires. Another question that arises is what will need to change in this model system of control once there is a shift from the use of algorithms that work based on pre-input criteria towards AI that is more self-learning, like the deep neural networks described in the background chapter of this paper.

5.4 Summary

There are many interpretations of what accountability and responsibility mean exactly. This can be attributed to the difference between usage in practice and on paper, as well as the difference between those using the terms. Legal practitioners, policymakers, private companies and organizations, and wider society have slightly differing understandings of what 'accountability' means or should mean. The classic legal frameworks through to establish responsibility for wrongdoing or mistakes are based on the ability to prove individual or group intent behind wrongdoing. However, the fragmentation of military technology development means that direct legal responsibility, accountability and eventually liability are difficult to establish in the diversified and often long life cycles of the elements that make up RAS from design and R&D all the way past deployment to decommissioning. While other bodies of law have been suggested in order to look at the way to legally address the advent of RAS, there is currently no body of law in place that fully suffices. This means that accountability should be addressed at an institutional level all throughout the life cycle.

In light of the increased privatization of military technologies, responsibility is fragmented across many actors. It is therefore crucial to ensure that actors' behavior—be they contractor, military or otherwise—can be accounted for and that ways to ensure and evaluate this are institutionalized in RAS' governance.

6 | Conclusion and Recommendations

Robotic and autonomous systems are the latest frontier in the competition for technological dominance in the military domain. While the delegation of tasks to machines is not a new phenomenon, recent advances in computation are enabling machines to carry out increasingly complex tasks. These range from autonomous air-to-air refueling and landing in an independently selected location, to smart swarms of ground, air and naval systems operating in sync through machine-to-machine communication. Alongside technology's fast-paced development, ethical norms evolve continuously. This requires governments and military forces to reflect on how ethical issues and norms of the society they operate in may affect the use of RAS. A remaining unsolved question is how to respond to adversaries using RAS to create a significant military advantage and escalation dominance.

This paper has identified four overarching ethical challenges arising from the use of RAS. First, the establishment of human control in increasingly autonomous systems, based on determining *how* and *where* in the life cycle and the observe-orient-decide-act (OODA) loop to maintain control over RAS, as well as *who* should do so over *what* functions of systems. The second challenge is the technical complexity of (semi-) autonomous systems, which leads to decreasing explainability and predictability of the system design, self-learning abilities and software updates. The third challenge is posed by limitations of human cognition, such as automation bias and complacency, as well as the anthropomorphizing of machines, arising in human–machine teaming. Fourth is the institutional risk management of the outsourcing of system design and manufacturing, and RAS interoperability with technologically advanced allied forces.

Considering the increasing proliferation of autonomous systems, including among adversaries, the RNLA should continue to experiment with systems that may enhance its portfolio, without losing sight of foundational ethical principles. In considering the implementation policy for RAS, the RNLA should seek to translate the discussion on meaningful human control into operational terms, such as by identifying controversial or high-risk machine functions, as presented in Chapter 3. This selective approach to establishing and maintaining human control presents a balance between ethical concerns and military objectives.

Meanwhile, the debate from Chapter 4 on the extent to which fully autonomous systems could be developed used in ways fully compliant with IHL and could thereby be in line with what society's idea of human dignity may be, is one that underpins the ethical dilemmas surrounding RAS. The break-down of RAS into life cycle, sub-system elements and the OODA loop, as presented in Chapter 3, is relevant again in addressing the challenge that RAS pose for traditional responsibility and accountability in the military, as well as within the broader national and international governance in which RAS are used, as is discussed in Chapter 5. Care should be taken to anticipate future risks of increased autonomy and to address the possibility of a resulting accountability gap ahead of time. New frameworks or a different use of existing bodies of law may be necessary, for example by considering institutionalizing a 'system of control' involving all relevant actors throughout the entire life cycle of RAS. The detailed example of a 'system of control' is one of the ways accountability can be thoroughly woven into the institutional handling of RAS, considering the many different stakeholders involved in RAS development. This can help guide decisions on RAS integration into the armed forces in a way that considers ethical standards and ensures moral responsibility at all stages of RAS development and use.

The paper provides the following concrete recommendations to the Government of the Netherlands, the Netherlands Ministry of Defence, the Royal Netherlands Army, and other governments and armed forces at strategic and operational levels:

Strategic

1. *Institutional development* – adapt internal processes, such as monitoring & evaluation with RAS life cycles, to better address (rapid) technological developments in relation to ethical issues. This means hiring or working closely with experts to guide use case development, testing with private contractors, and facilitating the introduction of RAS within the RNLA;
2. *Ethics by design* – there is a need to develop a set of guidelines for identifying use-cases, designing, validating, and manufacturing ethical RAS in line with the core principles of International Humanitarian Law and Article 36, rather than establishing ethical considerations only at the deployment stage;
3. *Testing* – in determining the appropriate environment for testing RAS, the RNLA may consider studying the testing approaches of other states and determine how to best emulate them while respecting core ethical principles;
4. *Contracting* – identify best practices for military–private sector cooperation in designing, manufacturing, maintaining, and operating military RAS, and delineate legal and moral responsibility for accidents, failures, malfunctions, or misuse of systems among the involved parties;
5. *Transparency* – communicate the Ministry of Defence and RNLA's research into and use of RAS to the public in order to inform and add nuance to the discussion on the value of RAS to the RNLA outside of the dominant 'killer robot' narrative;
6. *Accountability* – Ingrained throughout the entire life cycle of RAS should be an institutional approach to the roles and responsibilities of all actors involved. This helps ensure meaningful human control at all stages and creates a culture of shared accountability.

7. *Research* – continue to research the role of RAS in the military context, including but not limited to human–machine teaming, embedding of ethics in machines and contingency planning for facing adversaries using RAS based on lower ethical considerations, and focus on operationalizing these principles into practical applications.

Operational

1. *Human control* – considering the spectrum of autonomy, predetermine where, how and who maintains control over what functions of individual systems, as well as who is responsible for the initiation, use and shut down of systems prior to their deployment;
2. *Selective automation* – identify within what functions and why increasing automation and autonomy will benefit the military without eliciting major ethical concerns, e.g., movement controls and sensory controls;
3. *Interoperability* – push for standardization frameworks among technologically advanced allied forces on the training, deployment and operation of RAS in shared environments or during joint missions;
4. *Design process* – involve the end-users (i.e. operators and supervisors) in the use case development, design and testing phases to ensure the design of human-machine interfaces is suited to those using the systems;
5. *Opacity* – tackle the 'black box' nature of complex systems by developing traceability or logic flow processes to enable operators or supervisors to understand, explain and predict the operation of RAS;
6. *Training manuals* – in cooperation with contractors, issue training manuals for operators, supervisors and commanders of RAS for the initial use and after subsequent software updates that substantially alter the behavior or decision-making process of the system*;*
7. *Operation of RAS* – establish and delineate the different levels of freedom within the rules of engagement, based on the degree of autonomy of the system(s) under their command for military units deploying RAS;
8. *Human–machine teaming* – identify the limitations of human cognition in the oversight of RAS, develop understanding of its effect on human-machine teaming, and channel acquired knowledge in the RAS human interface design;
9. *Rules of engagement* – within the design and manufacturing process, seek to program fundamental rules of engagement (ROEs) with International Humanitarian Law principles embedded in system design, along with an open architecture to introduce mission-specific ROEs by mission command;
10. *Command responsibility* – predetermine command responsibility for the use of RAS for every individual deployment.

Bibliography

"Accountability." In *Business Dictionary*, n.d. http://www.businessdictionary.com/definition/accountability.html.

"AEGIS Weapon System." US Navy Fact File. US Navy, January 10, 2019. https://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=200&ct=2.

"Annex 3-60 – Targeting. Appendix A: Targeting and Legal Consideration. Basic Principles of the Law of War and Their Targeting Implications." U.S. Air Force Doctrine, March 15, 2019. https://www.doctrine.af.mil/Portals/61/documents/Annex_3-60/3-60-D33-Target-LOAC.pdf.

Arkin, Ronald C. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton F.L.: CRC Press Taylor & Francis Group, 2009.

———. "Lethal Autonomous Systems and the Plight of the Non-Combatant." *AISB Quarterly*, July 2013.

Asaro, Peter. "Jus Nascendi, Robotic Weapons and the Martens Clause." In *Robot Law*, by Ryan Calo, A. Michael Froomkin, and Ian Kerr, 2016. https://www.elgaronline.com/view/edcoll/9781783476725/9781783476725.00024.xml.

———. "The Liability Problem for Autonomous Artificial Agents." In *AAAI Spring Symposia*, 2016.

"Australia's System of Control and Applications for Autonomous Weapon Systems." Geneva, 2019. https://www.unog.ch/80256EDD006B8954/(httpAssets)/39A4B669B8AC2111C12583C1005F73CF/$file/CCW_GGE.1_2019_WP.2_final.pdf.

Banta, Benjamin R. "'The Sort of War They Deserve'? The Ethics of Emerging Air Power and the Debate over Warbots." *Journal of Military Ethics* 17, no. 2–3 (July 3, 2018): 156–71. https://doi.org/10.1080/15027570.2018.1551320.

Bendett, S. "In AI, Russia Is Hustling to Catch Up." *Defense One*, April 4, 2018. https://www.defenseone.com/ideas/2018/04/russia-races-forward-ai-development/147178/.

Bhana, Hemant. "By the Book - Good Written Guidance and Procedures Reduce Pilots' Automation Complacency." Aero Safety World, March 2010. https://flightsafety.org/wp-content/uploads/2016/11/asw_mar10_p47-51.pdf.

Bo, Marta, and Taylor Woodcock. "Blog: Lethal Autonomous Weapons, War Crimes, and the Convention on Conventional Weapons." *Asser Institute* (blog), May 28, 2019. https://www.asser.nl/about-the-institute/asser-today/blog-lethal-autonomous-weapons-war-crimes-and-the-convention-on-conventional-weapons/.

Boeing. "Maintenance Program Enhancements." *Boeing*, 2006. http://www.boeing.com/commercial/aeromagazine/articles/qtr_4_06/article_05_2.html.

Boogaard, Jeroen van den. "Proportionality and Autonomous Weapons Systems." Opinio Juris, 2016. http://opiniojuris.org/2016/03/23/proportionality-and-autonomous-weapons-systems/.

Boothby, William, ed. *New Technologies and the Law in War and Peace*. Cambridge University Press, 2018. https://www.cambridge.org/core/books/new-technologies-and-the-law-in-war-and-peace/AA47FC74ABEA568F6971E53EF906601C.

Bouvier, Antoine A. "International Humanitarian Law and the Law of Armed Conflict." Edited by Harvey J. Langholtz. Peace Operations Training Institute, 2012. http://cdn.peaceopstraining.org/course_promos/international_humanitarian_law/international_humanitarian_law_english.pdf.

Boyd, John. "The Essence of Winning and Losing." 1995. https://www.danford.net/boyd/essence.htm.

Boyle, Michael J. "The Legal and Ethical Implications of Drone Warfare." *The International Journal of Human Rights* 19, no. 2 (February 17, 2015): 105–26. https://doi.org/10.1080/13642987.2014.991210.

Boyle v United Techs. Corp. 487 U.S. 500, 510 (1988) (n.d.).

Brenton, Richard, and Eloi Bosse. "The Cognitive Costs and Benefits of Automation." In *RTO-MP-088*, 2002. https://www.researchgate.net/publication/235171183_The_Cognitive_Costs_and_Benefits_of_Automation.

Bryson, Joanna, Mihailis Diamantis, and Thomas Grant. "Of, For, and By the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence and Law* 25, no. 3 (September 30, 2017): 273–91.

Burrell, Jenna. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3, no. 1 (June 1, 2016): 2053951715622512. https://doi.org/10.1177/2053951715622512.

Calloway, Audra. "Army Wargames Shape the Future of Urban Warfare." U.S.Army, March 1, 2019. https://www.army.mil/article/215731/army_wargames_shape_the_future_of_urban_warfare.

Cassese, Antonio. "The Martens Clause: Half a Loaf or Simply Pie in the Sky?" *European Journal of International Law* 11, no. 1 (2000): 187–216.

Chairperson of the Informal Meeting of Experts. "Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)." United Nations Office at Geneva, 2016.

https://www.unog.ch/80256EDD006B8954/(httpAssets)/DDC13B243BA863E6C1257FDB00380A88/$file/ReportLAWS_2016_AdvancedVersion.pdf.

Charisi, Vicky, Louise Dennis, Michael Fisher, Robert Lieck, Andreas Matthias, Marija Slavkovik, Janina Loh, Alan Winfield, and Roman Yampolskiy. "Towards Moral Autonomous Systems," November 1, 2017. https://arxiv.org/pdf/1703.04741.pdf.

Chehtman, Alejandro. "New Technologies Symposium: Autonomous Weapons Systems– Why Keeping a 'Human On the Loop' Is Not Enough." *Opinio Juris* (blog), May 8, 2019. https://opiniojuris.org/2019/05/08/new-technologies-symposium-autonomous-weapons-systems-why-keeping-a-human-on-the-loop-is-not-enough/.

Chung, Timothy. "OFFensive Swarm-Enabled Tactics." Defense Advanced Research Projects Agency (DARPA). Accessed August 14, 2019. https://www.darpa.mil/program/offensive-swarm-enabled-tactics.

Clarke, Arthur C. *Profiles of the Future; an Inquiry into the Limits of the Possible*. New York: Harper & Row, 1973.

Cohen, Stanley. *States of Denial: Knowing about Atrocities and Suffering*. Polity, 2001.

Committee on Legal Affairs. "Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))." European Parliament, May 31, 2016. http://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect.

"Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field. Geneva, 12 August 1949." International Committee of the Red Cross (ICRC), 1949. https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/ART/365-570004?OpenDocument.

"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, 2019. https://www.unog.ch/80256EDD006B8954/(httpAssets)/39A4B669B8AC2111C12583C1005F73CF/$file/CCW_GGE.1_2019_WP.2_final.pdf.

Crawford, Kate, and Meredith Whittaker. "Artificial Intelligence Is Hard to See." Medium, September 11, 2016. https://medium.com/@katecrawford/artificial-intelligence-is-hard-to-see-a71e74f386db.

Crootof, Rebecca. "War Torts: Accountability for Autonomous Weapons." *University of Pennsylvania Law Review* 164 (May 2016): 56.

Cummings, Mary. "Automation Bias in Intelligent Time Critical Decision Support Systems." *American Institute of Aeronautics and Astronautics*, no. Session: IS-15: Human Interaction with Intelligent Systems (September 2004). https://doi.org/10.2514/6.2004-6313.

"Customary IHL - Practice Relating to Rule 14. Proportionality in Attack." ICRC IHL Database. Accessed June 25, 2019. https://ihl-databases.icrc.org/customary-ihl/eng/docs/v2_rul_rule14.

Danner, Allison and Martinez, Jenny, "Guilty Associations: Joint Criminal Enterprise, Command and Responsibility, and the Development of International Criminal Law." California Law Review 93, no. 1 (2005): 77-170.

Davison, Neil. "A Legal Perspective :Autonomous Weapon Systems under International Humanitarian Law." *UNODA Occasional Papers* 30 (2018).

———. "Autonomous Weapon Systems: An Ethical Basis for Human Control?" ICRC Humanitarian Law & Policy Blog, March 4, 2018. https://blogs.icrc.org/law-and-policy/2018/04/03/autonomous-weapon-systems-ethical-basis-human-control/.

Declaration Renouncing the Use, in Time of War, of Explosive Projectiles Under 400 Grammes Weight (1868). https://ihl-databases.icrc.org/ihl/full/declaration1868.

Deeks, Ashley, Noam Lubell, and Daragh Murray. "Machine Learning, Artificial Intelligence, and the Use of Force by States." *Journal of National Security Law & Policy* 10, no. 1 (2019): 1–25.

Development, Concepts and Doctrine Centre. *Human-Machine Teaming*. Joint Concept Note, 1/18. UK Ministry of Defence, 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/709359/20180517-concepts_uk_human_machine_teaming_jcn_1_18.pdf.

———. "Unmanned Aircraft Systems - Joint Doctrine Publication 0-30.2." UK Ministry of Defence, 2017. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/673940/doctrine_uk_uas_jdp_0_30_2.pdf.

Dickinson, Laura. "Contract as a Tool for Regulating Private Military Companies." In *From Mercenaries to Market*, by Simon Chesterman and Chia Lehnardt. Oxford University Press, 2007. https://doi.org/10.1093/acprof:oso/9780199228485.001.0001.

———. *Outsourcing War and Peace*. Yale University Press, 2011. https://yalebooks.yale.edu/book/9780300144864/outsourcing-war-and-peace.

Dickinson, Laura A. "Drones, Automated Weapons, and Private Military Contractors." In *New Technologies for Human Rights Law and Practice*, edited by Molly K. Land and Jay D. Aronson, 93–124. Cambridge University Press, 2018. https://doi.org/10.1017/9781316838952.005.

"Dilbert at War." *The Economist*, June 23, 2014. https://www.economist.com/united-states/2014/06/23/dilbert-at-war.

Docherty, Bonnie. "Banning 'Killer Robots': The Legal Obligations of the Martens Clause." Arms Control Association, January 10, 2018. https://www.armscontrol.org/act/2018-10/features/remarks-banning-%E2%80%98killer-robots%E2%80%99-legal-obligations-martens-clause.

———. "Heed the Call: A Moral and Legal Imperative to Ban Killer Robots." USA: Human Rights Watch, August 21, 2018. https://www.hrw.org/report/2018/08/21/heed-call/moral-and-legal-imperative-ban-killer-robots.

———. "Losing Humanity: The Case against Killer Robots." Human Rights Watch, 2012.

Ekelhof, Merel. "Autonome Wapensystemen: Wat We Moeten Weten over de Toepassing van Het Humanitair Oorlogsrecht En de Menselijke Rol in Militaire Besluitvorming." *Ars Aequi*, March 2018, 193–202.

———. "Autonomous Weapons: Operationalizing Meaningful Human Control." Humanitarian Law & Policy Blog, August 15, 2018. https://blogs.icrc.org/law-and-policy/2018/08/15/autonomous-weapons-operationalizing-meaningful-human-control/.

———. "Lifting the Fog of Targeting: 'Autonomous Weapons' and Human Control through the Lens of Military Targeting." *Naval War College Review* 71, no. 3 (2018): 61–94.

Etzioni, Amitai, and Oren Etzioni. "Pros and Cons of Autonomous Weapons Systems." *Military Review* 97, no. 3 (June 2017): 72–81.

European Commission. "Commission Regulation (EU) No 1321/2014 of 26 November 2014 on the Continuing Airworthiness of Aircraft and Aeronautical Products, Parts and Appliances, and on the Approval of Organisations and Personnel Involved in These Tasks." Office Journal of the European Union, December 17, 2014. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014R1321&from=EN.

European Group on Ethics in Science and New Technologies (EGE). *Statement on Artificial Intelligence, Robotics and "autonomous" Systems*. Luxembourg: European Commission Directorate-General for Research and Innovation, 2018.

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

Feickert, Andrew, Lawrence Kapp, Jennifer K Elsea, and Laurie A Harris. "U.S. Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress." U.S. Congressional Research Service, November 20, 2018. https://digital.library.unt.edu/ark:/67531/metadc1442984/.

Floridi, Luciano. "What the Near Future of Artificial Intelligence Could Be." *Philosophy & Technology* 32, no. 1 (March 1, 2019): 1–15. https://doi.org/10.1007/s13347-019-00345-y.

Floridi, Luciano, Jessica Morley, Libby Kinsey, and Anat Elhalal. "From What to How - An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices," 2019. https://www.academia.edu/39135750/From_What_to_How-_An_Overview_of_AI_Ethics_Tools_Methods_and_Research_to_Translate_Principles_into_Practices.

Folkman, Joseph. "The '8 Great' Accountability Skills For Business Success." *Forbes*, November 14, 2014. https://www.forbes.com/sites/joefolkman/2014/11/14/how-do-you-score-the-8-great-accountability-skills-for-business-success/#74093a803c11.

Galand, Alexandre, Emilie Hunter, and Ilia Utmelidze. "International Criminal Law Guidelines: Command Responsibility." Case Matrix Network, January 2016. https://www.legal-tools.org/doc/7441a2/pdf/.

Giesen, Ivo, and François Kristen. "Liability, Responsibility and Accountability: Crossing Borders." *Utrecht Law Review* 10, no. 3 (June 2014): 1–13.

Giger, Jean-Christophe, Nuno Piçarra, Patrícia Alves-Oliveira, Raquel Oliveira, and Patricia Arriaga. "Humanization of Robots: Is It Really Such a Good Idea?" *Human Behaviour & Emergining Technologies*, 2019, 111–23.

Gilbert, David. "Russian Weapons Maker Kalashnikov Developing Killer AI Robots." *Vice News*, July 13, 2017. https://news.vice.com/en_us/article/vbzq8y/russian-weapons-maker-kalashnikov-developing-killer-ai-robots.

Gray, H.M., K. Gray, and D.M. Wegner. "Dimensions of Mind Perception." *Science* 315, no. 5812 (2007): 619.

Grishenko, Nikolai. "Российский Подводный Робот Выполнил Боевую Задачу в Сирии." *Российская Газета (RG.RU)*, February 22, 2018. https://rg.ru/2018/02/22/rossijskij-podvodnyj-robot-vypolnil-boevuiu-zadachu-v-sirii.html.

Gunkel, David. “Other Things: AI, Robots and Society.” In *A Networked Self and Human Augmentics, Artificial Intelligence, Sentience*, 1st ed., 216. Routledge, 2018. https://www.taylorfrancis.com/books/e/9781315202082.

Han, Yi, Benjamin I. P. Rubinstein, Tamas Abraham, Tansu Alpcan, Olivier De Vel, Sarah Erfani, David Hubczenko, Christopher Leckie, and Paul Montague. “Reinforcement Learning for Autonomous Defence in Software-Defined Networking.” In *Decision and Game Theory for Security*, edited by Linda Bushnell, Radha Poovendran, and Tamer Başar, 11199:145–65. Springer International Publishing, 2018. https://doi.org/10.1007/978-3-030-01554-1_9.

Hawkins, Andrew. “Everything You Need to Know about the Boeing 737 Max Airplane Crashes.” *The Verge*, March 22, 2019. https://www.theverge.com/2019/3/22/18275736/boeing-737-max-plane-crashes-grounded-problems-info-details-explained-reasons.

Henderson-Sellers, Brian, and Julian M. Edwards. “The Object-Oriented Systems Life Cycle.” *Communications of the ACM* 33, no. 9 (September 1990): 142–159. https://doi.org/10.1145/83880.84529.

“Hero-400EC Extended-Range Loitering System.” Air Force Technology, n.d. https://www.airforce-technology.com/projects/hero-400ec-extended-range-loitering-system/.

Hoek, Freek, Cor van Montfort, and Cees Vermeer. “Enhancing Public Accountability in the Netherlands.” *OECD Journal on Budgeting* 5, no. 2 (2005): 70–86.

Hoijtink, Marijn, and Matthias Leese. *Technology and Agency in International Relations*. 1st ed. Routledge, 2019. https://www.taylorfrancis.com/books/e/9780429463143.

Horowitz, Jonathan. “Joint Blog Series: Precautionary Measures in Urban Warfare: A Commander’s Obligation to Obtain Information.” Humanitarian Law & Policy Blog, October 1, 2019. https://blogs.icrc.org/law-and-policy/2019/01/10/joint-blog-series-precautionary-measures-urban-warfare-commander-s-obligation-obtain-information/.

Horowitz, Michael C., and Paul Scharre. “Meaningful Human Control in Weapon Systems: A Primer.” Working paper. Project on Ethical Autonomy. Center for a New American Security (CNAS), March 2015. https://www.files.ethz.ch/isn/189786/Ethical_Autonomy_Working_Paper_031315.pdf.

Hsu, Jeremy. “Real Soldiers Love Their Robot Brethren.” *Live Science*, May 21, 2009. https://www.livescience.com/5432-real-soldiers-love-robot-brethren.html.

Hughes, Joshua. “No, Autonomous Weapon Systems Are Not Unlawful under the Martens Clause.” Medium, August 21, 2018. https://medium.com/@jghughes1991/no-autonomous-weapon-systems-are-not-unlawful-under-the-martens-clause-2653d18790e9.

International Committee of the Red Cross ICRC. “Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?” Geneva, Switzerland: International Committee of the Red Cross (ICRC), 2018. https://www.icrc.org/en/download/file/69961/icrc_ethics_and_autonomous_weapon_systems_report_3_april_2018.pdf.

International Law Commission. “Responsibility of States for Internationally Wrongful Acts.” United Nations, 2001. http://legal.un.org/ilc/texts/instruments/english/draft_articles/9_6_2001.pdf.

JASON, The MITRE Corporation. “Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD,” January 2017. https://fas.org/irp/agency/dod/jason/ai-dod.pdf.

Johansson, Linda. “Ethical Aspects of Military Maritime and Aerial Autonomous Systems.” *Journal of Military Ethics* 17, no. 2–3 (July 3, 2018): 140–55. https://doi.org/10.1080/15027570.2018.1552512.

Jones, Bruce, Charles T Call, Daniel Toubolets, and Jason Fritz. “Managing the New Threat Landscape: Adapting the Tools of International Peace and Security.” *Brookings Institute*, Foreign Policy at Brookings, September 2018, 23.

Kania, Elsa. “China’s Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems.” Lawfare, April 17, 2018. https://www.lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems.

———. “China’s Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems.” Lawfare, April 17, 2018. https://www.lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems.

———. “The Critical Human Element in the Machine Age of Warfare.” Foundatio. *Courier* (blog), n.d. https://www.stanleyfoundation.org/articles.cfm?id=867&title=A-Happy-Place--to-Be-a-Cow.

Kania, Elsa B. “Battlefield Singularity: Artificial Intelligence, Military Revolution, and China’s Future Military Power.” Washington, D.C.: Center for a New American Security (CNAS), November 2017. https://s3.amazonaws.com/files.cnas.org/documents/Battlefield-Singularity-November-2017.pdf?mtime=20171129235805.

Khan, Azmat and Anand Gopal, “The Uncounted.” *The New York Times*, November 16, 2017.

"Killer Robots - Learn." Campaign to Stop Killer Robots. Accessed March 7, 2019. https://www.stopkillerrobots.org/learn/.

"Killer Robots and the Concept of Meaningful Human Control." Human Rights Watch, 2016. https://www.hrw.org/sites/default/files/supporting_resources/robots_meaningful_human_control_final.pdf.

"Killer Robots Fail Key Moral, Legal Test: Principles and Public Conscience Call for Preemptive Ban." *Human Rights Watch* (blog), August 21, 2018. https://www.hrw.org/news/2018/08/21/killer-robots-fail-key-moral-legal-test.

Klare, Michael T. "Autonomous Weapons Systems and the Laws of War." Arms Control Association, March 2019. https://www.armscontrol.org/act/2019-03/features/autonomous-weapons-systems-laws-war.

Koohi v. United States, 976 F.2d 1328, 1336–37 (9th Cir. 1992) (n.d.).

Kool, R.S.B. "(Crime) Victims' Compensation: The Emergence of Convergence." *Utrecht Law Review* 10, no. 3 (June 2014): 14–26.

Krishman, Armin. *Killer Robots: Legality and Ethicality of Autonomous Weapons*. Routledge, 2016.

Kumar, Padma, Sharafat Hussain, Alexis Espiritu, Lj Marzo, and Rosa Rakavono. "Algorithms, Flowcharts, Data Types and Pseudocode." Accessed August 14, 2019. https://www.academia.edu/34581869/2._ALGORITHMS_FLOWCHARTS_DATA_TYPES_AND_PSEUDOCODE_2.1_ALGORITHMS.

Legality of the Threat or Use of Nuclear Weapons (Advisory Opinion of 8 July 1996), 1996 Reports 226 (ICJ 1996).

Lehman, M. M. "Programs, Life Cycles, and Laws of Software Evolution." *Proceedings of the IEEE* 68, no. 9 (September 1980): 1060–76. https://doi.org/10.1109/PROC.1980.11805.

Leveringhaus, Alex. "Autonomous Weapons Mini-Series: Distance, Weapons Technology and Humanity in Armed Conflict." Humanitarian Law & Policy Blog, October 6, 2017. https://blogs.icrc.org/law-and-policy/2017/10/06/distance-weapons-technology-and-humanity-in-armed-conflict/.

Lewis, Larry. "AI and Autonomy in War: Understanding and Mitigating Risks." CNA Center for Autonomy and AI, n.d. https://www.cna.org/CNA_files/PDF/Understanding-Risks.pdf.

Lijn, Jaïr van der, and Stefanie Ros. "Peacekeeping Contributor Profile: The Netherlands." Providing for Peacekeeping, January 2014. http://www.providingforpeacekeeping.org/2014/04/08/contributor-profile-the-netherlands/.

Lin, Patrick, George Bekey, and Keith Abney. "Autonomous Military Robotics: Risk, Ethics, and Design." Ethics + Emerging Sciences Group at California Polytechnic State University, December 20, 2008. https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1001&context=phil_fac.

———. "Robots in War: Issues of Risk and Ethics." In *Ethics and Robotics*, edited by R. Capurro and M. Nagenborg, 49–67. AKA Verlag Heidelberg, 2009. https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1010&context=phil_fac.

Liu, Hin-Yan. "Contract Law as Cover: Curtailing the Scope of Private Military and Security Contractor Responsibilities." *The Ashgate Research Companion to Outsourcing Security: The Role of the Market in 21st Century Warfare (Joakim Berndtsson and Christopher Kinsey Eds.)*, 2016. https://www.academia.edu/15301427/Contract_Law_as_Cover_Curtailing_the_Scope_of_Private_Military_and_Security_Contractor_Responsibilities.

Long, Drake. "China Releases Video of 56-Boat Drone Swarm near Hong Kong." *TheDefensePost*, June 2, 2018. https://thedefensepost.com/2018/06/02/china-56-boat-drone-swarm-hong-kong/.

Lu, Denise, Allison McCann, Jin Wu, and K. K. Rebecca Lai. "From 8,600 Flights to Zero: Grounding the Boeing 737 Max 8." *The New York Times*, March 13, 2019. https://www.nytimes.com/interactive/2019/03/11/world/boeing-737-max-which-airlines.html.

Luban, David. [italics] Legal Modernism: Law, Meaning, and Violence. *University of Michigan Press*, 1994.

Malik, Swati. "Autonomous Weapon Systems: The Possibility and Probability of Accountability." *Wisconsin International Law Journal* 35, no. 3 (n.d.): 34.

Marchant, Gary, Ronald C. Arkin, Braden Allenby, Edward T. Barrett, Jason Borenstein, Lyn Gaudet, Orde Kittrie, et al. "International Governance of Autonomous Military Robots." *Columbia Science & Technology Law Review* 272 (2011).

Marra, William C., and Sonia K. Mcneil. "Understanding 'The Loop': Regulating the Next Generation of War Machines." *Harvard Journal of Law & Public Policy* 36, no. 3 (May 2013): 1139–85.

Matthias, Andreas. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6, no. 3 (September 2004): 175–83.

McIntyre, Alison. "Doctrine of Double Effect." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2019. Metaphysics Research Lab, Stanford University, 2019. https://plato.stanford.edu/archives/spr2019/entries/double-effect/.

McLean, Wayne. “Drones Are Cheap, Soldiers Are Not: A Cost-Benefit Analysis of War.” The Conversation. Accessed June 28, 2019. http://theconversation.com/drones-are-cheap-soldiers-are-not-a-cost-benefit-analysis-of-war-27924.

Melzer, Nils. “Keeping the Balance between Military Necessity and Humanity – a Response to Four Critiques of the ICRC’s Interpretive Guidance on the Notion of Direct Participation in Hostilities.” *New York University Journal of International Law and Politics* 42 (2010): 831–916.

Miller, Christopher A., and Raja Parasuraman. “Designing for Flexible Interaction Between Humans and Automation: Delegation Interfaces for Supervisory Control.” *Human Factors* 49, no. 1 (February 1, 2007): 57–75. https://doi.org/10.1518/001872007779598037.

“Milrem Robotics Delivered Two THeMIS UGVs to the Dutch Army.” Private. Milrem Robotics, May 28, 2019. https://milremrobotics.com/milrem-robotics-delivered-two-themis-ugvs-to-the-dutch-army/.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. “The Ethics of Algorithms: Mapping the Debate.” *Big Data & Society* 3, no. 2 (December 1, 2016): 2053951716679679. https://doi.org/10.1177/2053951716679679.

Mohanty, Bedavyasa. “Lethal Autonomous Dragon: China’s Approach to Artificial Intelligence Weapons.” *ORF* (blog), November 15, 2017. https://www.orfonline.org/expert-speak/lethal-autonomous-weapons-dragon-china-approach-artificial-intelligence/.

Mucic et al., “Celebici”, No. IT-96-21-T (ICTY November 16, 1998).

Musgrave, Shawn. “Inside ‘Liberty City,’ Homeland Security’s Site for Testing Urban Drones.” Vice Motherboard, June 19, 2015. https://www.vice.com/en_us/article/wnj9nq/inside-liberty-city-homeland-securitys-site-for-testing-urban-drones.

Netherlands Advisory Council on International Affairs. “Autonomous Weapon Systems: The Need for Meaningful Human Control.” Netherlands Advisory Council on International Affairs, October 2015. https://aiv-advies.nl/8gr#government-responses.

Parasuraman, Raja, and Dietrich Manzey. “Complacency and Bias in Human Use of Automation: An Attentional Integration.” *Human Factors* 52, no. 3 (October 2010): 381–410. https://doi.org/10.1177/0018720810376055.

Pastor, E, C Barrado, P Royo, J Lopez, and E Santamaria. “An Open Architecture for the Integration of UAV Civil Applications, Aerial Vehicles.” In *Aerial Vehicles*, 511–36. IntechOpen, 2009. https://www.intechopen.com/books/aerial_vehicles/an_open_architecture_for_the_integration_of_uav_civil_applications.

Preece, Alun. “Asking ‘Why’ in AI: Explainability of Intelligent Systems – Perspectives and Challenges.” *Intelligent Systems in Accounting, Finance and Management* 25, no. 2 (April 19, 2018). https://doi.org/10.1002/isaf.1422.

“Preparing for More Urban Warfare.” *The Economist*, January 25, 2018. https://www.economist.com/special-report/2018/01/25/preparing-for-more-urban-warfare.

Press, Michael. “Of Robots and Rules: Autonomous Weapon Systems in the Law of Armed Conflict.” *Georgetown Journal of International Law* 48 (2017): 1337–66.

Prosecutor v Kavishema, No. ICTR-95-1-T (ICTR May 21, 1999).

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

“Reframing Autonomous Weapons Systems.” In *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems (A/IS)*. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, n.d. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_reframing_autonomous_weapons_v2.pdf.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” February 16, 2016. http://arxiv.org/abs/1602.04938.

Robert, Lionel. “The Growing Problem of Humanizing Robots.” *International Robotics & Automation Journal* 3, no. 1 (2017). https://medcraveonline.com/IRATJ/IRATJ-03-00043.pdf.

Roff, Heather M., and David Danks. “‘Trust but Verify’: The Difficulty of Trusting Autonomous Weapons Systems.” *Journal of Military Ethics* 17, no. 1 (January 2, 2018): 2–20. https://doi.org/10.1080/15027570.2018.1481907.

“RQ-11 Raven Unmanned Aerial Vehicle.” Army Technology, n.d. https://www.army-technology.com/projects/rq11-raven/.

Sandoz, Yves, Christophe Swinarski, and Bruno Zimmermann. “Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949.” International Commission of the Red Cross, 1987. https://perma.cc/5XKM-QQYV.

Santoni de Sio, Filippo, and Jeroen van den Hoven. “Meaningful Human Control over Autonomous Systems: A Philosophical Account.” *Frontiers in Robotics and AI* 5, no. 15 (2018). https://doi.org/10.3389/frobt.2018.00015.

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. WW Norton & Co, 2018.

Scharre, Paul, and Michael Horowitz. "Meaningful Human Control in Weapon Systems: A Primer." Center for New American Security, March 2015. https://www.files.ethz.ch/isn/189786/Ethical_Autonomy_Working_Paper_031315.pdf.

Schmitt, Michael, and Jeffrey Thurnher. "'Out of the Loop': Autonomous Weapon Systems and the Law of Armed Conflict." *Harvard National Security Journal* 4, no. 231 (2013). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2212188.

Schwarz, Elke. "Intelligent Weapons Systems and Meaningful Human Control: An Uneasy Alliance," 2019. Working Paper - available from author.

———. "The (Im)Possibility of Meaningful Human Control for Lethal Autonomous Weapon Systems." International Committee of the Red Cross. *Humanitarian Law & Policy* (blog), August 29, 2018. https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/.

Scott, Andrew, Jose Solorzano, Jonathan Moyer, and Barry Hughes. "Modeling Artificial Intelligence and Exploring Its Impact." Frederick S. Pardee Center for International Futures Josef Korbel School of International Studies University of Denver, May 2017. https://pardee.du.edu/sites/default/files/ArtificialIntelligenceIntegratedPaper_V6_clean.pdf.

Sharkey, Noel. "Guidelines for the Human Control of Weapons Systems." ICRAC, April 2018. https://www.icrac.net/icrac-working-paper-3-ccw-gge-april-2018-guidelines-for-the-human-control-of-weapons-systems/.

———. "Killer Robots From Russia Without Love." *Forbes*, November 28, 2018. https://www.forbes.com/sites/noelsharkey/2018/11/28/killer-robots-from-russia-without-love/.

Sharkey, Noel E. "The Evitability of Autonomous Robot Warfare." *International Review of the Red Cross*, Comments and Opinions, 94, no. 886 (Summer 2012): 787–99. https://doi.org/10.1017/S1816383112000732.

Shilo, Liron. "Speaking of Responsibility: Autonomous Weapon Systems, State and Individual Responsibility." Georgetown Law, Institute for Technology, Law & Policy. *Georgetown Tech* (blog), n.d. https://www.georgetowntech.org/blogfulltext/2017/5/1.

Sparrow, Rob. "Ethics as a Source of Law: The Martens Clause and Autonomous Weapons." Humanitarian Law & Policy Blog, November 14, 2017. https://blogs.icrc.org/law-and-policy/2017/11/14/ethics-source-law-martens-clause-autonomous-weapons/.

Sparrow, Robert. "Building a Better Warbot: Ethical Issues in the Design of Unmanned Systems for Military Applications." *Science and Engineering Ethics* 15, no. 2 (2009): 169–187. https://doi.org/10.1007/s11948-008-9107-0.

Spiegeleire, Stephan De, Matthijs Maas, and Tim Sweijs. *Artificial Intelligence and the Future of Defense*. The Hague, The Netherlands: The Hague Centre For Strategic Studies, 2017. https://hcss.nl/sites/default/files/files/reports/Artificial%20Intelligence%20and%20the%20Future%20of%20Defense.pdf.

"Statement of the Head of the Russian Federation Delegation, Director of the Department for Nonproliferation and Arms Control of the Russian Ministry for Foreign Affairs V.Yermakov at the Meeting of the State-Parties of the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons on Item 7 of the Agenda 'General Exchange of Views', Geneva, November 21, 2018." The Ministry of Foreign Affairs of the Russian Federation, November 22, 2018. http://www.mid.ru/main_en/-/asset_publisher/G51iJnfMMNKX/content/id/3415655.

Stewart, Emily. "The Boeing 737 Max 8 Crashes and Controversy, Explained." *Vox*, March 13, 2019. https://www.vox.com/2019/3/12/18262359/boeing-737-max-controversy-faa-trump.

Strawser, Bradley Jay. "Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles." *Journal of Military Ethics* 9, no. 4 (December 1, 2010): 342–68. https://doi.org/10.1080/15027570.2010.536403.

Suchman, Lucy. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge University Press, 2006. https://www.cambridge.org/core/books/humanmachine-reconfigurations/9D53E602BA9BB5209271460F92D00EFE.

"The Montreux Document - On Pertinent International Legal Obligations and Good Practices for States Related to Operations of Private Military and Security Companies during Armed Conflict." *International Committee of the Red Cross*, 2009, 48.

"The Position Paper Submitted by the Chinese Delegation to CCW 5th Review Conference." Geneva: United Nations Office at Geneva, n.d. https://www.unog.ch/80256EDD006B8954/(httpAssets)/DD1551E60648CEBBC125808A005954FA/$file/China%27s+Position+Paper.pdf.

Torossian, Bianca, Frank Bekkers, Tim Sweijs, Michel Roelen, Alen Hristov, and Salma Atalla. "Paper on the Military Applicability of Robotic and Autonomous Systems." The Hague Centre for Strategic Studies, Paper forthcoming - available from authors 2019.

Travis, Gregory. "How the Boeing 737 Max Disaster Looks to a Software Developer." *IEEE Spectrum*, April 18, 2019. https://spectrum.ieee.org/aerospace/aviation/how-the-boeing-737-max-disaster-looks-to-a-software-developer.

United Nations, Department of Economic and Social Affairs, Population Division. "World Urbanization Prospects: The 2018 Revision," 2018.

United Nations Institute for Disarmament Research (UNIDIR. "The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence." United Nations, 2018. http://www.unidir.ch/files/publications/pdfs/the-weaponization-of-increasingly-autonomous-technologies-artificial-intelligence-en-700.pdf.

———. "The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward." United Nations, 2014. http://www.unidir.ch/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf.

"Unmanned Systems Integrated Roadmap FY 2011-2036." US Department of Defense, 2011. https://fas.org/irp/program/collect/usroadmap2011.pdf.

US Department of Defense. "Directive 3009.09." US Government, November 21, 2012. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf.

Verdiesen, Ilse. "Agency Perception and Moral Values Related to Autonomous Weapons: An Empirical Study Using the Value-Sensitive Design Approach. Masters Thesis." TU Delft, 2017.

Vincent, James. "Giving Robots 'Personhood' Is Actually about Making Corporations Accountable." The Verge, January 19, 2017. https://www.theverge.com/2017/1/19/14322334/robot-electronic-persons-eu-report-liability-civil-suits.

Weisgerber, Marcus. "What's in the House NDAA?; Pentagon's 3D-Mapping Service; New Marine One, Weed Whacker; and More." Defense One, May 10, 2018. https://www.defenseone.com/business/2018/05/global-business-brief-may-10-2018/148116/.

Worcester, Maxim. "Autonomous Warfare – A Revolution in Military Affairs." *ISPSW Strategy Series: Focus on Defense and International Security*, no. 340 (April 2015): 6.

Wu, Xiang-Hu, Ming-Cheng Qu, Zhi-Qiang Liu, and Jian-Zhong Li. "Research and Application of Code Automatic Generation Algorithm Based on Structured Flowchart." *Journal of Software Engineering and Applications* 4 (2011): 534–45. https://doi.org/10.4236/jsea.2011.49062.

Appendix A - Expert Session Scenarios[248]

Scenario 1. 'Killerbot'

ISIS has been defeated in Syria by rooting out the last members in Baghouz on the Syrian-Iraqi border. Unfortunately, ISIS ideas have not been eradicated, and an influential leader Abu Bakr-al Baghdadi (ABaB) still prophets his vision of an Islamic Caliphate through his re-occurring presence in the media. ABaB has gone underground, but it is clear that his extreme ideas are gaining traction and groups of young Jihadis are currently being mobilized to plan and conduct terror attacks in their domiciles in Western Europe. It is essential for domestic safety in our country that ABaB is silenced as soon as possible. A major intelligence operation has been conducted and as a result, deployed HUMINT units learned that he is in rural Syria, in a 50km^2 area of mountainous terrain with a myriad of tunnels. The tunnels are not charted and are likely booby-trapped. As such, it will be very difficult to thoroughly comb through the area.

Major Pavlov Strolsky, the local commander of the Russian Spetznatz-unit who was responsible for finding ABaB, has to plan an operation to eliminate ABaB's role as the Jihadi leader. Because of time sensitivity, short exposure time of ABaB, and the need to prevent him from fleeing again, there is no time to discuss alternatives between major Strolsky and his headquarters. It is important to note that communications from within the caves to the outside HQ is impossible due to the iron-ore rich stone of the mountains.

The nations with troops in the area (Russia, Turkey, Syria) are reluctant to conduct searches (Russia and Syria) or are prohibited by their own government (Turkey) through a lack of Rules of Engagement (ROE) concerning these kinds of operations. A solution would be sending out a ground drone loaded with facial recognition software and armed with lethal capabilities. Two months ago, the Russians brought the 'Gusenichnyy' (Crawler) ground drone system into theater for operational evaluation and testing. Civilian (Russian) personnel from Kamaz, the producer of the Gusenichnyy system, assist the military in handling the drone. Although the crawler is a rather small object (60cm high and 35kg), it can move over rocks easily, and is silent, stealthy, and lethal. Experiments in a safe environment have proven that the facial recognition software has 99.95% accuracy and has the ability to self-learn in order to further minimize errors. ABaB's facial features have already been loaded into the 'Gusenichnyy' by civilian engineer Pjotr Pekar of Kamaz. As instructed by the

[248] **Disclaimer**: Scenarios 1-4 are entirely fictional but use the names of real companies, states, and locations only for the purpose of a more in-depth simulation. All the names of individuals are fictional. The content of the scenarios is not intended to discredit any companies or states.

management of Kamaz, Pekar urges Major Stroslky to deploy the crawler for this specific mission in order to get the first real test data on performance, in terms of recognition tasks and lethal tasks.

It is likely that there is only one chance to find and stop ABaB, so if the crawler is deployed, it has to be set on fully autonomous mode. That means that it fires lethally when the software recognizes ABaB. It is Major Strolsky's decision to either send the Gusenichnyy or send his highly trained men into the caves and risk losing them.

Discuss the decision-making considerations, both in ethical and legal sense, which Major Strolsky should or could be contemplating within this case.

Scenario 2. 'Testbed'

The year is 2021 and the Netherlands government is struggling with the last pieces of legislation concerning autonomous armed drones in combat. To fill this gap, it wants to enhance testing, in order to prepare legislation for military use outside of the Netherlands. This knowledge is essential for building legislation relevant to all military activities (including unmanned and lethal) and to show how the military will operate with these systems under water, at the surface or in air and space. Theoretical experiments have been conducted already, but the last piece of information concerning real-life flying and drone-weapon separation (with live weapons) still has to be acquired.

The Commander of the Royal Netherlands Air Force, LtGen Frits Elands, proposes a test over the Vliehors shooting and bombing range on the isle of Vlieland. His plan involves ceasing air operations and training in the area for one day. He proposes the drones take off from Leeuwarden Air Base (which also tested the Reaper MQ9 drone that entered service in 2020) in order to minimize flying time over inhabited areas. The autonomous software is capable of finding the specifically designed target at the range. The drone is programmed to only fire a missile at the predesigned target (which is unique in shape, colors, and pattern). Stringent safety margins, including geo-fencing, are in place.

If and when the test is successful, legislation can be concluded and various unmanned systems owned by the Dutch military will be ready for deployment.

The US urges the Netherlands to take an expeditious approach, as they want more European involvement in their mission in the Sinai (Operation Vanguard against some virulent, widespread, covert operating ISIL units), where a few Dutch troops are currently present alongside the American contingent.

The Netherlands ground drones, developed by the Dutch firm VDL, fill in urgent operational capability gaps in the US operation. VDL is willing to assist in all technical matters to conduct the test, as they foresee great US interest in buying their drone. This would lead to a €2.5 billion deal and 1,500 jobs in the Netherlands for the next 10 years. The pressure on the Dutch government from industry and foreign allies to conduct the testing, finish legislation, and deploy the drones is high.

Discuss the required legal framework and the ethical/moral implications of this testing. What are the consequences for quick deployment of Dutch drones and how relevant is the positive spin-off for the Dutch economy?

Scenario 3. 'Defend'

Netherlands troops conduct a peace-enforcement operation in the Central African Republic (CAR). Their Camp 'Hoogeveen' houses a total of 1,800 Dutch soldiers from the HQ, the Helicopter Task Force, the Special Forces patrol, armored infantry, artillery and support units. No civilians work here, nor is the camp close to built-up areas with local CAR-nationals. The Chief of Defense of the Netherlands, General van de Putte, is currently visiting with the Danish Chief of Defense (General Norrebro), as in a week's time the Danish will start their operational hand-over from the Netherlands. In one month, the Netherlands will end this mission after four years of fighting local insurgents. Eight Dutch soldiers have died in the past 2.5 years. Intelligence reports indicate that this high-level visit will be used by insurgents to conduct attacks on Camp Hoogeveen as a means of maximum exposure in the media.

Because of these VIP visits, the Danes have provided ground surveillance and protection through their armed drone unit 'Thor', which consists of autonomous ground systems that can work in swarms. The Danes have made serious breakthroughs in research concerning drones with high-energy weapons, which are lethal. Because of the terrain, distances, and employed tactics by insurgents, the ground-based drones work on autonomous setting. This means that random search patterns will be 'walked' by the drones and their weapons can be fired (based on previously set algorithms and taking the ROEs into full consideration). Unclear to the Dutch staff is what will trigger the drones to use their deadly lasers. Confirmed, however, is that under no circumstance a laser will be used toward or within the boundaries of the camp.

Another means of protection would be stationing almost all soldiers on a rotational schedule outside the Camp's perimeter. This seems to be of high risk, due to the advanced night-fighting capabilities the insurgents have shown to possess in combination with the level of exhaustion the Dutch troops suffer. The Dutch troops will be rotating out in only a few weeks time after enduring six months of extreme hardship.

It is up to the operational commander of the camp, Colonel Peter van Ellekom to take a decision on defensive measures.

Discuss the decision making, ethics and legality of using a defensive tactic with drones when the 'only' possible victims of an adversary attack are trained Dutch soldiers.

Scenario 4. 'Back to basic'

Netherlands forces are in Lithuania conducting their Enhanced Forward Presence mission (EFP), together with German, English, and Polish troops. They defend the city of Narva (56,350 inhabitants), bordering on the Russian border. Tensions are high, as the Russian S400 system in Kaliningrad has been active for two weeks already and NATO jets performing Baltic Air Policing missions have been 'painted' not only by the surveillance radar, but also by the tracking and fire-control radars of SA7 and SA21 air defense systems. NATO thinks it unwise to deploy manned fighters, due to risks of losing a fighter to a Russian rocket and fears of the escalation that would follow such an attack. In order to de-escalate, NATO deploys some drones instead of armed fighters to keep watch over the ground troops in the city. Western intelligence organizations receive insider information that Russia is planning a devastating air attack on Narva.

In order to defend troops and civilians, defensive measures have to be taken, otherwise a large proportion of city inhabitants and the deployed EFP troops will fall victim to Russia's lethal air attacks within the highly populous medieval town center. NATO's UAVs are not armed (so as not to further escalate), but they are in direct communication with unmanned ground systems equipped with anti-aircraft artillery.

The ground-based drone with these weapons is the so-called 'Umbrella' system. 'Umbrellas' are positioned along the border with Russia and throughout the town on the roofs of high buildings. Due to extremely short reaction times, 'Umbrella' has to be set on autonomous operations. This means that once opposing aircraft will cross the border from Russia to NATO territory, Umbrella will fire upon them. NATO has communicated this to the Russian high command in St. Petersburg and to the Russian President.

Discuss how such a defense strategy would fit within NATO's show of resilience and coherence in defending NATO territory vis-a-vis article 5. Discuss how ethical concerns interact with the prospect of preventing the killing of thousands of civilians.

Appendix B - Expert Session Scenario Summaries

Scenario 1. 'Killerbot'

Killerbot is evidently the most controversial scenario, particularly in terms of human dignity and human control. Trust was a central issue within human control, as concerns were raised over how a system that was only tested in a "safe" (rather than a combat) environment be trusted to complete the mission. Further concern was the inability to communicate with the crawler or abort/alter the course of action if necessary. Questions arose over meaningful human control (MHC), particularly in terms of what qualifies as a threshold for MHC and whether programming of RAS is enough human control. With regard to human dignity, concern was raised over the inability of the RAS to understand situational nuance, e.g., the target surrendering, use of human shields and presence of non-combatants in the environment. The inaccuracy of 0.05% was seen as a major issue in light of civilians present in the operational environment. As the system had not been tested in a combat environment by the military, the overall performance of the system in line with Article 36 of the Geneva Convention Protocol I was seen as an issue. One group was willing to accept civilian casualties under the Doctrine of Double Effect[249] seeing as the target is a high value target. However, the threshold for the acceptable number of civilian casualties was not determined. As operators of the RAS, responsibility was laid with the Russians (unspecified at what level), but with concerns that the private sector would remain unaccountable for the extent to which it is involved in the development. Moreover, the issue of trust arose, whereby the text was suggestive of the Russian military placing trust in the private sector engineers with little to no experience with the RAS and without regard for Article 36. The self-learning of the system was also perceived as controversial, as this also reduces human control, predictability, and thus, has significant implications for responsibility. Overall, conclusions of whether to send the crawler in were mixed. Some participants said to unequivocally "go for it", while others said the system was doomed to fail and would not meet norms and rules of military engagement. Most participants fell somewhere in the middle, with sharp concerns for human control and human dignity contrasting their recognition that the use of RAS was likely to be the most effective and least risky action given the almost impossible circumstances of the scenario.

[249] The Doctrine of Double Effect "is often invoked to explain the permissibility of an action that causes a serious harm, such as the death of a human being, as a side effect of promoting some good end. According to the principle of double effect, sometimes it is permissible to cause a harm as a side effect (or "double effect") of bringing about a good result even though it would not be permissible to cause such a harm as a means to bringing about the same good end." See McIntyre, "Doctrine of Double Effect."

Scenario 2. ‘Testbed’

The scenario discussions focused primarily on human control and human dignity. The controversy was centered around human dignity, based on the idea that an individual exercise in a controlled environment is not representative of a real combat scenario. As a result, there is insufficient evidence that the UAV would execute ‘ethical’ decisions in actual combat. Similarly, limited testing highlighted a lack of human control, both in manufacturing and operations. In manufacturing, the principal-agent problem between the private sector and the military resulting from pressure to deploy RAS quickly is likely to result in the deployment of premature systems. In operations, the lack of familiarity of the operator with the UAV could reduce the explainability of the system, in turn reducing operator’s control, with implications for responsibility. Human control and responsibility is further at risk due to software patches and self-learning of the RAS, meaning over time explainability would reduce further. As a result, testing will need to be carried out after every software update, as it could alter the system’s behavior. More testing is necessary to develop the operator’s familiarity with the system to limit the possibility for the operator to be sheltered from responsibility by claiming that the machine acted unpredictably. As a result, one group argued that testing should focus on confirming operational parameters and the functioning of the system, rather than simulating operational behavior, which would still be far from an actual operational environment. Furthermore, test environments are not capable of testing the chain of command, and hence responsibility, as the outcomes are predetermined and hence, the operators do not experience the pressure of a real life scenario.

Solutions proposed to the challenges were based on an understanding that further testing would be necessary, but the pursuit of RAS will continue, driven by competition from adversaries and the faster pace of technological development in the private sector. One group suggested that testing can also provide insight into the ways other actors will use RAS, even if the Dutch government does not allow the deployment of certain RAS in the future. One solution was to focus on ‘development’ legislation rather than ‘deployment’ legislation, arguing that this would enable further testing and more extensive control over the private sector. One of the groups also noted the difficulty of legislating unpredictable behavior. In the case that testing is limited, RAS should only be allowed to carry out missions highly comparable to those it has executed in a controlled environment.

Scenario 3. ‘Defend’

With the scenario as it was written, there was a consensus that the lethal autonomous system could not be used. This was largely down to the fact that the Netherlands troops were not familiar with the Danish system in question and they did not know what would trigger the lethal ray the system's swarmed robots could shoot. All groups concluded that for the use of this, or any autonomous system in the military, it is crucial that the troops using it fully understand the system they are using and can predict the outcome of whatever situation or environment the system is used in. What the necessary level of (training) experience with a system is, or how much information the commander needs to receive from a party delivering the system remained a point of discussion, and is clearly a topic that requires further research. Some participants were of the opinion that Dutch troops would need to be well acquainted with and have trained themselves for quite some years with RAS before being able to ethically and predictably deploy them. Others thought that when collaborating with trusted partners it could be enough to receive a certain amount of information on the systems used, including perhaps testing reports, relevant indicators, ROEs, and more such data. Difference in levels of not only confidence but also trust in fellow soldiers as opposed to RAS is partly down to the extent to which one can understand the reasoning process of each. Even if a human acts irrationally, there is a certain reasoning process behind the actions that can be explained afterward. As close as possible a level of understanding is needed of the reasoning process that guides the actions of RAS.

Another question that came to the foreground was whether the full level of autonomy was even necessary in this situation. Many found that, if given the option, the best solution would be the teaming of a semi-autonomous swarm and human response in case of threat or confrontation. Reasons to reconsider this unwillingness to opt for full autonomy came in the form of situations of pressure, such as where there is severe time pressure or an environment in which humans could perform well enough, e.g., difficult terrain or severe exhaustion. Choices are made on the basis of (imminent) risk analysis and the question of how necessary a certain level of autonomy is to prevent the further endangering of human lives. Even so, it was said that such systems should have the option to perform in various modes and not as autonomous systems only, and that there should be serious consideration of in which cases or settings lethality is an option.

Some groups already went in the direction of solutions, with the most concrete being the development of an international, standardized system for the various types of RAS that partners may use. Once the use of RAS for various purposes in international contexts is more normalized, it will be important that partners can be quickly made aware of which system it is they are dealing with, what this type of system's outcomes are based on, how it has been tested, how the system can be used, and more such crucial aspects of confidence in decisions on the (joint) use or avoidance of RAS in certain military contexts.

Scenario 4. 'Back to Basic'

Scenario 4 was generally straightforward and uncontroversial. There was an overall consensus among the groups that the umbrella system should be deployed, with one group going as far as to state that "it would be unethical not to deploy the RAS". The main reason for the conviction was that the system is defensive by nature and that the actions were explicitly communicated to the Russian higher command. The system is likely to serve as a better deterrence tool as it is in autonomous mode, thus limiting the strategic choices of the adversary. Some noted that it would be important for the adversary to know the demonstrated potential of the RAS to limit the temptation to test the system. Ethical issues were almost untouched, with groups rather side-tracking to operational and political issues, notably a point flagged by one of the groups that the Russians may perceive the deployment of RAS on the border as an escalation. Frequent comparisons were drawn to the Aegis system, the Patriot missiles and the Israeli Iron Dome, so the 'availability heuristic' was quite prevalent across the groups. Primary controversies in ethics were false positives, i.e. the shooting down of non-military aircraft (e.g. USS Vincennes incident) or situations like the Russia-Turkey dispute over the shooting down of the Russian aircraft over the Turkey-Syria border, with the idea that rapid decision-making by the RAS can lead to unnecessary escalation if the threat is not explicitly demonstrated. Particularly due to the proximity of the border, the area of the umbrella system's operation should be clearly defined to avoid takedowns of aircraft still within Russian airspace. Some disagreements were evident in terms of human control, with arguments between autonomous mode vs keeping humans in the loop. Little discussion on NATO Article 5, with most agreeing that Article 4 was more fitting for the scenario, with the possibility of an escalation to Article 5 should an attack occur. To address proportionality concerns, one group suggested illuminating/marking a target first and firing only in case of violation of warnings.

Appendix C - List of Abbreviations

AACUS – Autonomous Aerial Cargo/Utility System

ACTUV – Anti-Submarine Warfare Continuous Trail Unmanned Vessel

AGI – Artificial General Intelligence

AI – Artificial Intelligence

ASW – Anti-submarine warfare

(L)AWS – (Lethal) Autonomous Weapon Systems

CCW – Convention on Certain Conventional Weapons

DL – Deep learning

DNN – Deep neural network

EFP – Enhanced Forward Presence

GGE – Group of Governmental Experts

HARM – High-speed anti-radiation missile

HCSS – *The Hague* Centre for Strategic Studies

ICJ – International Court of Justice

ICL – International Criminal Law

ICRC – International Committee of the Red Cross

IHL – International Humanitarian Law

ILC – International Law Commission

LOAC – Law of Armed Conflict

MHC – Meaningful human control

ML – Machine learning

NATO – North Atlantic Treaty Organization

OODA – Observe-orient-decide-act

PMSC – Private military and security contractor

R&D – Research and development

RAS – Robotic and autonomous system

RNLA – Royal Netherlands Army

ROE – Rules of engagement

TNO – Netherlands Organisation for Applied Scientific Research

UAV – Unmanned aerial vehicle

UGV – Unmanned ground vehicle