1

Autonomous Weapons Systems and International Relations

Throughout the twentieth century, the story of modern, industrial-scale warfare undertaken by highly developed states revolved around increasing the physical distance between soldiers and their enemies or targets: from air campaigns in the Second World War to the development of cruise missiles during the Cold War; from networked warfare in the Persian Gulf War's Operation Desert Storm to remote warfare via drones. This technology-mediated process changed the character of warfare significantly, as Peter Singer summarised in the context of drone operators: 'Going to war has meant the same thing for 5,000 years. Now going to war means sitting in front of a computer screen for 12 hours' (Helmore 2009). In fact, 'direct human involvement has been reducing in modern warfare over time' (Lele 2019, 51).

Autonomous weapons systems not only continue this trajectory of the *physical* absence of humans from the battlefield but also introduce their *psychological* absence 'in the sense that computers will determine when and against whom force is released' (Heyns 2016a, 4). As this broad understanding implies, autonomous features in weapons systems can take many different forms: we find them in loitering munitions, in aerial combat vehicles, in stationary sentries, in counter-drone systems, in surface vehicles, and in ground vehicles (Boulanin and Verbruggen 2017).

Even military technology that has been around for much longer, sometimes decades, such as air defence systems and active protection systems, has automated or autonomous targeting qualities. Table 1.1 provides an overview of some weapons systems in current use or development that include automated or autonomous features in the use phase of the targeting process. In other words, they have such features in their *critical* functions. This does not mean that these systems are used without human control. In fact, we typically find that systems with autonomous features can be operated in distinct modes with different levels of human control. To give an example, the US-employed Aegis air defence system 'has four modes, ranging from "semiautomatic", where a human operator controls decisions regarding the use of lethal force, to "casualty", which assumes that the human operators are incapacitated and therefore permits the system to use defensive force independently' (Crootof 2015, 1858–59).

But what matters is that these systems can, in principle while not (currently) in practice, detect, engage, and attack targets autonomously, without further human intervention. Weapons systems with autonomous features are not a topic of the distant future – they are already in use (see also Scharre 2018, 4).

These diverse systems are somewhat uneasily captured by the catch-all category of autonomous weapons systems (AWS), because they weaponise artificial intelligence (AI) and apply this in varying combat situations. They signal the potential absence of *immediate* human decision-making on lethal force and the incremental loss of so-called meaningful human control, a concept that has become a central focus of the transnational debate on lethal autonomous weapons systems (LAWS) at the Convention on Certain Conventional Weapons (CCW), as we will demonstrate in more detail in the course of this book. While lethality is a potential outcome of AWS usage, what is problematic about integrating autonomous features applies in general to 'acting with the intent to cause physical harm, i.e. violence' (Asaro 2019, 541; see also Rosert and Sauer 2020, 14; Heyns 2016b, 355; Crootof 2015, 1836). We therefore use the more general term 'AWS' throughout the book and only refer to LAWS when speaking to the transnational debate at the CCW, as the discussion there is specifically focused on *lethal* autonomous weapons systems.

The first section introduces AWS and their role by way of conceptual definitions as well as the implications for international relations in two steps. First, we introduce AWS and discuss the competing definitions of autonomy and human–machine interaction. We also clarify the associated conceptual terminology, such as AI and machine learning. Second, we shed light on the ongoing transnational, political debate on AWS at the United Nations (UN) in Geneva under the auspices of the UN-CCW. Third, we review the academic literature on AWS, characterising it as chiefly interested in their (potential) legality, ethical challenges attached to their usage, and options for their (legal) regulation. Thus, this chapter provides an empirical basis for the remainder of the book, introducing readers to the core fault lines in the debate on AWS.

Table 1.1 Selected weapons systems with automated or autonomous features in their critical functions.

Active protection systems	Iron Curtain (US) *in operation*
	Korean Active Protection System (Republic of Korea) *in operation*
	Trophy/ASPRO-A/Windbreaker (Israel) *in operation*
Anti-personnel sentry weapons	Samsung SGR-A1 (Republic of Korea) *in operation*
	Sentry Tech (Israel) *in operation*
Air defence systems	Iron Dome (Israel) *in operation*
	MIM-104 Patriot (US) *in operation*
	Phalanx (US) *in operation*
Combat air vehicles	X-47B (US) *tech demonstrator*
	Taranis (UK) *tech demonstrator*
Ground vehicles	Uran-9 (Russia) *in operation*
	Robotic Technology Demonstrator (US) *tech demonstrator*
Counter-drone systems	HEL effector (Germany) *in operation*
	Drone Dome (Israel) *in operation*
	Silent Archer (US) *in operation*
Guided munitions	Dual-Mode Brimstone (UK) *in operation*
	Mark 60 CAPTOR (US) *in operation*
Loitering munitions	Harpy, Harop (Israel) *in operation*
	KARGU-2 (Turkey) *in operation*
	FireFly (Israel) *in operation*
Surface vehicles	Sea Hunter II (US) *development completed*
	Protector USV (Israel) *in operation*

Sources: Boulanin and Verbruggen (2017), Boulanin (2016), Roff (2016), and Holland (2019).

AUTONOMOUS FEATURES, ARTIFICIAL INTELLIGENCE, AND AUTONOMOUS WEAPONS SYSTEMS

Autonomous features are what make the development and deployment of AWS significant. Yet, what autonomy is or should refer to is a matter of contestation and remains 'poorly understood' (Haas and Fischer 2017, 285). Likewise, the discourse about AWS frequently uses 'automation' and 'autonomy' interchangeably. It is easy to get bogged down in the ongoing debate about definitions – in fact, this lack of consensus has hampered discussions among states at the CCW (Ekelhof 2017). We should also recognise that how participants at the CCW define autonomy is deeply political, as it 'affects what technologies or practices they identify as problematic and their orientation toward a potential regulatory response' (Brehm 2017, 13).

Contributors to the debate on AWS – be they states, institutions, or defence manufacturers – invariably have a stake in defining autonomy in ways that advance their interests. To illustrate, actors may use the terms 'automated' or 'highly automated' rather than 'autonomous' in referring to a weapons system's critical functions because they imply a greater level of human control. Likewise, they may add stringent requirements to any definition of autonomy in order to avoid the regulation or condemnation of systems currently in development (Moyes 2019). Defence companies often 'play up the sophistication and autonomy of their products in marketing, and downplay them when scrutinised by international bodies such as the United Nations (Artificial Intelligence Committee 2018, 26).

To navigate these ambiguities and contextualise our subsequent analysis, we only provide basic, workable definitions of autonomy and automation. Autonomy is a relative concept and can be broadly defined as the 'ability of a machine to perform a task without human input' (Scharre and Horowitz 2015, 5). On this basis, an autonomous system is one that 'once activated, can perform some tasks or functions on its own' (Boulanin and Verbruggen 2017, 5). Automation is a term that overlaps with these understandings and is often used synonymously with it. The difference between the two is not always clear (Hagström 2016, 23).

Basic definitions found in robotics offer some distinctions. According to robot ethicist Alan Winfield, automation means 'running through a fixed preprogrammed sequence of actions', while autonomy means that 'actions are determined by its sensory inputs, rather than where it is in a preprogrammed sequence' (Winfield 2012, 12). To further unpack this, while automated systems follow clearly defined, *deterministic* 'if–then' rules, autonomous systems *select* probabilistically defined best courses of action – this makes the output of automated systems predictable, while conversely implying that autonomous systems will produce uncertain outputs rather than consistently producing the same results (Cummings 2018, 8). As Heyns notes, autonomous systems 'are unpredictable in the sense that it is impossible to foresee all of the potential situations that they will encounter during their programming' (Heyns 2016b, 356; see also Crootof 2016, 1350).

Autonomous systems combine software and hardware components that fulfil three key functions: *perception* via different sensors (e.g. electro-optical, infrared, radar, sonar) that allow the system to sense its environment; 'decision'/ *cognition*, through processors and software that assess 'collected data about the environment and plans courses of action' (IPRAW 2017a, 18); and *actuation*, allowing the system to physically respond in line with its planned courses of action (e.g. a motor triggering movement, or weapons release) (Welsh 2015; IPRAW 2017a, 18–19).

It is clear that basic levels of autonomy are comparatively easy to achieve. Indeed, a robotic vacuum cleaner, such as the popular Roomba, is a good example for such autonomy. It can navigate (i.e. change direction), make decisions, and take actions (i.e. clean particular spots on the floor more or less thoroughly) on the basis of its sensor inputs rather than where it is in a preprogrammed sequence. This makes it 'autonomous but not very smart' (Winfield 2012, 13; see also Roff 2015a).

Autonomy therefore does not necessarily imply a high level of 'sophistication' or intelligence. As Sharkey highlights, 'the autonomous robots being discussed for military applications are closer in operation to your washing machine than a science fiction Terminator' (Sharkey 2010, 376). Consequently, prominent AI researchers such as Toby Walsh argue that giving autonomy to such stupid, yet already available systems[1] is what constitutes the primary problem when considering lethal autonomous weapons systems (T. Walsh 2018, 189).

AI: a basic understanding

We can shine more light on this discussion about smart and stupid weapons by defining what we mean by AI in more detail. Defined in simple terms, AI is the 'attempt to make computers do the kinds of things that humans and animals do' (Boden 2017). In other words, 'AI is the capability of a computer system to perform tasks that normally require human intelligence, such as visual perception, speech recognition and decision-making' (Cummings 2018, 7), while the constituent components of intelligence also remain a matter of debate (see T. Walsh 2018, 61–81). There are a wide range of computational techniques summarised under the term 'AI' that are based on 'applications of mathematical logic, [and] advanced statistics' (IPRAW 2017b, 9) – any autonomous system is likely to require different techniques at different levels.

We can further distinguish between weak (also called 'narrow') AI and strong AI. Weak AI refers to applications that are capable of executing a single, particular task within a narrow domain in a way that 'equals or exceeds "human" capabilities' (T. Walsh 2018, 126). This goal has been reached in games such as chess, and to some extent in (much-hyped) speech and facial recognition (Dickson 2017). Context-specificity is an important feature of narrow AI, as even small changes to context and task specifications prevent the AI system from 'retain[ing] its level of intelligence' (Goertzel 2014, 1). This is because of the way much of narrow (weak) AI, which is typically based on variations of machine learning algorithms, learns: rather than being able to learn across problems the way humans do, '[m]achine learning algorithms tend to have to start again from scratch' (T. Walsh 2018, 94). This failure to generalise makes algorithms inherently brittle (Cummings 2018, 12–13). We will return to some of these specifics around machine learning shortly.

In contrast to weak AI, strong AI refers to 'machines that will be minds', including such essential features as 'self-awareness, sentience, emotion, and morality' (T. Walsh 2018, 126). AI researchers also refer to 'artificial general intelligence', which is supposed to approximate human-level intelligence: it comes with the interactive capability to self-adapt to different circumstances by generalising knowledge across tasks and contexts (Goertzel 2014, 2; Dickson 2017). This is also referred to simply as having common sense. In other words, machines with artificial general intelligence would have 'the ability to work on any problem that humans can do, at or above the level of humans' (T. Walsh 2018, 128).

As noted, not only has weak AI been achieved in some fields, in stark contrast to strong AI, but most research of the AI community focuses on weak AI applications (Roberts 2016; T. Walsh 2018, 127). Artificial general intelligence has been an unachieved research goal for decades, and its challenges are often highlighted by the so-called Moravec Paradox, which continues to influence the field: 'it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility' (Moravec 1988, 15). This makes the hard tasks easy to automate while making the easy tasks hard to automate.

Finally, there are also ideas about an artificial super intelligence that could exceed human-level intelligence, leading to the so-called technological singularity, an idea that has been around since at least the 1950s (T. Walsh 2018, 163–4). Most recently, Kurzweil, a prominent futurist, has popularised this line of thinking in associating the coming singularity primarily with the pace at which 'human-created technology' and, in particular, AI is growing (Kurzweil 2006, 7). As a result of this development, 'information-based technologies will encompass all human knowledge and proficiency, ultimately including the pattern-recognition powers, problem-solving skills, and emotional and moral intelligence of the human brain itself' (Kurzweil 2006, 8). At the point of the singularity, 'we build a machine that is able to redesign itself to improve its intelligence – at which point its intelligence starts to grow exponentially, quickly exceeding human intelligence by orders of magnitude' (T. Walsh 2018, 164). The philosopher Nick Bostrom has expressed similar ideas about the ongoing intelligence explosion that we find ourselves in, which is associated with coming machine superintelligence (Bostrom 2016).

Yet, as Walsh argues, 'the singularity is … an idea mostly believed by people *not* working in artificial intelligence' (T. Walsh 2017b; our emphasis). Instead, there continue to be significant doubts about the prospects of linear expectations of growth and progress as well as fundamental limits to innovation associated with any scientific field (IPRAW 2017b; LeVine 2017). Indeed, throughout its history, the AI research community has gone through various intermittent phases of progress euphoria and significant setbacks (T. Walsh 2017a). Walsh summarises a range of further arguments that fundamentally question this idea of a runaway, uncontrollable development of AI (T. Walsh 2017a, 166–78).

Rather than entering into (potentially obstructive) speculations about the likely developmental trajectory of AI, Roberts (2016) offers a succinct summary of the state of the art in AI research: 'Most of the work in the field for the past 40 years has focused on refining [artificial narrow intelligence] and better incorporating it into the human realm, taking advantage of what computers do well (sorting through massive amounts of data) and combining it with what humans are good at (using experience and intuition)'.

This leads us directly to consider various complex forms of human–machine interaction that will most likely be based on different forms of narrow AI and have the greatest potential to pose challenges.

Here, we should briefly engage with machine learning algorithms that are at the heart of (most) current applications in the area of narrow AI. This 'involves programming computers to teach themselves from data rather than instructing them to perform certain tasks in certain ways' (Buchanan and Miller 2017, 5). Machine learning revolves around producing *probabilistic* outputs that not only enable forms of diagnosis and description but are also increasingly used for prediction and prescription based on having identified features and patterns in data.

It is useful to differentiate between supervised and unsupervised learning. In supervised learning, various types of learning algorithms (e.g. support vector machines, decision trees, Bayesian networks) attempt to predict what connects input and output by using a labelled training data set (i.e. combining training examples with correct outputs). The goal is to find patterns within the data, allowing the learning algorithm to connect the correct input and the correct output. This is referred to as supervised learning for two reasons. First, 'each piece of data given to the algorithm also contains the correct answer about the characteristic of interest, such as whether an email is spam or not, so that the algorithm can learn from past data and test itself by making predictions' (Buchanan and Miller 2017, 6). In other words, all training data used is correctly labelled, i.e. 'this is a cat', 'this is a dog', etc. Second, humans can correct algorithmic outputs in real time during the learning phase. Supervised learning is conducted until the machine learning algorithm reaches a certain, reliable performance level (Brownlee 2016; IPRAW 2017b, 10–11). Conceptually, this is also connected to deep learning via neural networks that work through interconnections at multiple levels simultaneously. Supervised learning accounts for around 90% of machine-learning algorithms[2] and thus, perhaps unsurprisingly, most of the recent, well-publicised advances in AI have been in supervised learning (T. Walsh 2018, 95).

By contrast, in unsupervised learning, learning algorithms are fed with unlabelled input data that does not, therefore, come with correct outputs. In contrast to supervised learning, algorithms are charged with finding potentially interesting patterns, called clusters, within the data on their own (Brownlee 2016); 'they receive no input from the user of algorithm as to what the categories are or where the boundaries or lines might lie in the data' (IPRAW 2017b, 10). Unsupervised learning is considered to be beneficial in unstructured situations when 'there is not a clear outcome of interest about which to make a prediction or assessment' (Buchanan and Miller 2017, 8).

As this brief summary demonstrates, all machine-learning solutions are, by definition, data hungry and data dependent. This is typically characterised as machine learning's major problem because labelled data is not readily available, as 'in many application domains … collecting labels requires too much time and effort' (T. Walsh 2018, 95). Further, machine-learning solutions that are the outcome of supervised learning only work reliably with data that has been gained in the exact same manner as their training data sets:[3] 'the performance of the algorithm on further data, or real-world data, depends on how representative the training and test data sets are of the datasets in the application domain' (IPRAW 2017b, 11). To illustrate this with an example, autonomous cars that were trained with data gained in specific weather conditions and surroundings in California cannot be assumed to function in safe and reliable ways in other countries or regions: in fact, studies show that even small alterations to the training environment, such as changing weather and road conditions, make it hard to ensure safe and predictable driving behaviour by autonomous cars (Himmelreich 2018).

Apart from this reliance on particular types of training data, machine learning algorithms are faced with another, well-known fundamental problem that makes their integration into critical applications highly contestable from a safety and trust standpoint: they are a closed or black box. This means that they 'cannot meaningfully explain their decisions, why a particular input gives a certain output' nor can they 'guarantee certain behaviours' (T. Walsh 2018, 92).

These insights bring us to a wider set of practical questions: while civilian research advances in dual-use technologies such as AI are clearly relevant for autonomous features in weapons systems, we should not assume automatic spillover (Verbruggen 2019). Civilian applications differ contextually in important ways from military applications: for example, the unstructured environments of urban battlefields within which an autonomous ground vehicle would have to operate are significantly more challenging to navigate than the comparatively structured environment of highways that autonomous vehicles are currently trained for. And even here, as we have learned above, training data does not equal training data.

Typically, both civilian advances in AI and their direct connection to weaponised AI are often over-hyped. This is clearly visible in transnational discourse on AWS: at UN-CCW meetings in 2017 and 2018, everyone was buzzing about the rapid advances of Deep Mind's AlphaGo Zero, a Go-playing AI, towards 'superhuman' competency, as if this were a game-changing moment for the UN-CCW debate on the military applications of AI (Suchman 2018). In fact, the comparatively late timing of AlphaGo Zero's supposed breakthrough has rather surprised AI researchers, given that machines enjoy a home advantage in this tiny field. Games have long been a popular testing ground for AI because they offer access to a 'simple, idealized world' (T. Walsh 2018, 114): weak AI's superhuman capacity to number crunch makes it perfectly suited to finding rules-based solutions for one clearly contained goal in a constrained environment that can easily be iterated multiple times (Heath 2017).

Defining autonomy along a spectrum

For our purposes, we follow the definition of autonomy provided by researchers at the Stockholm Peace Research Institute (SIPRI): 'the ability of a machine to execute a task, or tasks, without input, using interactions of computer programming with the environment' (Boulanin and Verbruggen 2017, 5).

Autonomy should be thought of broadly as a 'relative independence' (Lele 2019, 55) and, importantly, as referring to particular features and functions of autonomous weapons systems, rather than to the system as a whole. This is an important distinction because some weapons systems differ significantly in their make-up of autonomous features: these can, for example, relate to their mobility, i.e. the capacity to direct their own motion, or health management, i.e. securing their functioning (Boulanin and Verbruggen 2017, 21–32). In connecting autonomy to core functions of weapons systems, such as trigger, targeting, navigation, and mobility, we can evaluate the extent to which a weapons system operates autonomously to different degrees (Roff 2015a). The crucial aspects of autonomy relate to autonomous features in trigger and targeting, that is in the *critical* functions of weapons systems to 'target select (search for, detect, identify, track or select) and attack (use force against, neutralize, damage or destroy)' (Davison 2017, 5).

Table 1.2 Spectrum of autonomy.

Remote controlled	Automated features	Autonomous features	Fully autonomous systems
Complex human–machine interaction			

Therefore, while the topic is often framed using the language of AWS as if only one, clear version of autonomy exists, what we see is an inclusion of autonomous features along a spectrum of autonomy (table 1.2).

At the one end of the spectrum, we find remote controlled systems where humans remain in manual control of the targeting functions, such as drones. Such systems require human input for executing their tasks. At the other end are what is often referred to as fully autonomous systems (see Heyns 2016a, 6). Here, humans are no longer involved in the specific use decisions. These are instead administered by the system, which operates completely on its own. But we see most significant developments in the middle of the spectrum: a zone that we have coined complex human–machine interaction. Here, systems exhibit automated and autonomous features and operate under the supervision of a human. This supervision differs in quality, depending on the range and type of tasks 'performed' via automated and autonomous features. We explore what this implies for *meaningful* human control in more detail in chapter 5.

The inclusion of more automated and autonomous features in the critical functions of weapons systems is likely to see humans move further and further away from *immediate* decision-making on using force. Often, this is highlighted by the image of the control loop, referring to human control in the so-called 'use phase', that is, in specific targeting situations (Human Rights Watch 2012, 2; J. Williams 2015, 183). Typically referred to as the orient, observe, decide, act (OODA) loop,[4] this image helps to visualise the relationship between the human and the system in specific situations when targets are selected and engaged, rather than in earlier phases of the targeting process, e.g. strategic planning (Burt 2018, 11). In 'in-the-loop' systems, humans actively participate in selecting specific targets and making decisions to use force. By contrast, in 'on-the-loop' systems, the role of the human operator is significantly reduced: they monitor system actions and can intervene when necessary, but ultimately only *react* to targets suggested by the program.

As this demonstrates, understanding the autonomy of weapons systems (almost always) involves various forms of human–machine interaction and the extent to which machine autonomy may (or does) undermine human autonomy. We will therefore consider human–machine interaction more closely, alongside the evolving concept of meaningful human control.

MEANINGFUL HUMAN CONTROL AND THE TARGETING PROCESS

Originally coined by the non-governmental organisation (NGO) Article 36 (Article 36 2013a), there are different understandings of what meaningful human control implies (see also chapter 5). Many states and other actors consider the application of violent force without any human control as unacceptable and morally reprehensible.

The term has gained significant currency in the transnational debate on LAWS (see section 1.3), yet it can refer to hugely different aspects. Sharkey's five levels of human supervisory control, for example, range from humans deliberating about specific targets before initiating an attack at the highest level, via humans choosing from a list of targets suggested by a program, to programs selecting targets and allocating humans a time-restricted veto at the lowest level (N. Sharkey 2016, 34–37). Brehm (2017, 8) offers a helpful indication of the contours of meaningful human control from a legal perspective: 'the requirement of meaningful human control over AWS would seem to entail that human agents involved in the use of an AWS have the opportunity and capacity to assess compliance with applicable legal norms and to take all legally required steps to respect and ensure respect for the law, including preventive and remedial measures'.

Further, even if we only concentrate on meaningful human control with regard to the critical weapons functions of selecting and attacking, targeting itself is a complex, multi-dimensional process in military terms (Ekelhof 2018). Following, for example, the North Atlantic Treaty Organization's doctrinal documentation, the (joint) targeting cycle goes through six phases: objectives/guidance; target development; capabilities analysis; commander's decision and assignment; mission planning and force execution; and assessment (see figure 1.1).

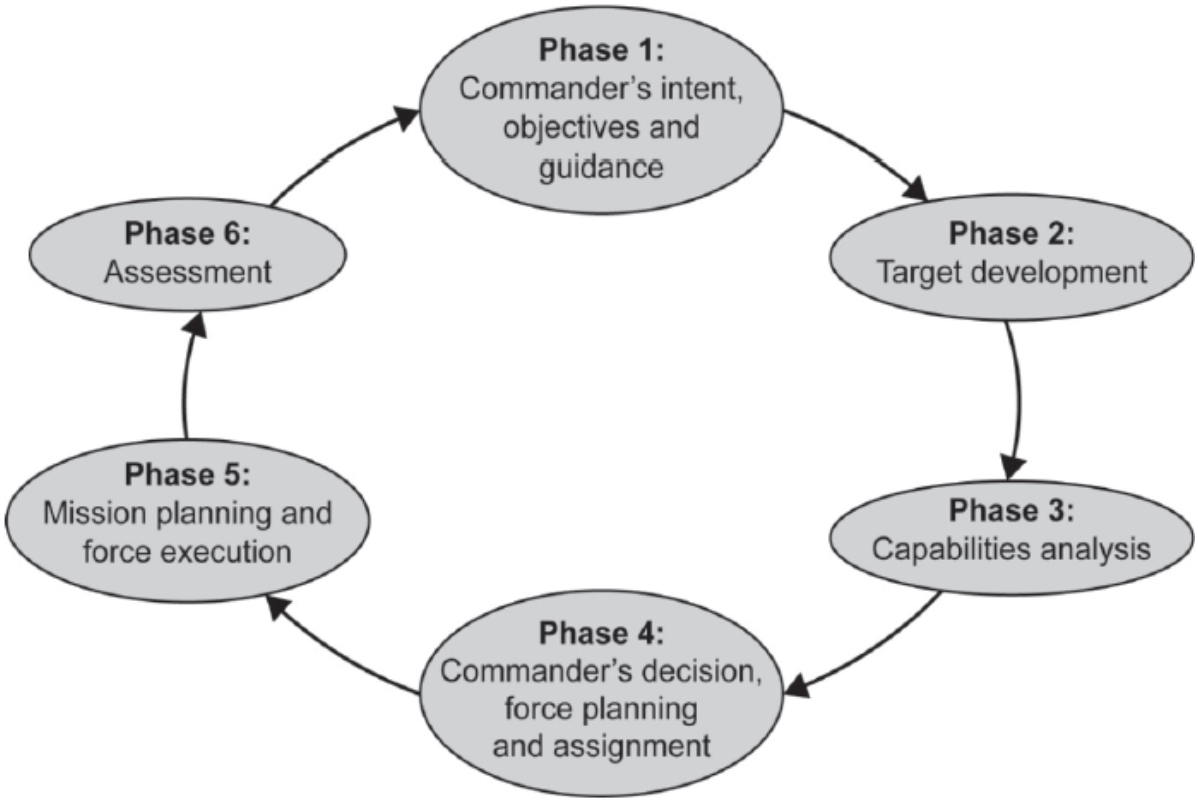

Figure 1.1 NATO's Joint Targeting Cycle.

Importantly, this conceptualisation encompasses 'deliberate planning phases of the targeting process' (Ekelhof 2018, 29, phases 1–4 and 6), rather than just their execution (phase 5). This more holistic approach to targeting indicates that, while some critical functions of weapons systems are not necessarily directly connected to the kinetic use of force, they can still contribute to target development.

This is an important consideration when thinking about how meaningful human control can be usefully defined in operational terms. To give an example: the United States has tested a machine-learning algorithm (the ill-named Skynet, another reference familiar to those acquainted with the *Terminator* franchise) to develop potential targets for drone operations from mobile phone metadata in Pakistan. While these pattern-identifying functions do not stand in direct connection to target engagement and attack, they are firmly, and importantly, part of the targeting cycle in terms of target selection and should therefore be included when we talk about critical functions of weapons systems.

Operational advantages and disadvantages of reducing human control

Retaining whatever form of *meaningful* human control is likely to be challenging in the light of operational considerations of effectiveness/efficiency, that is, identifying the human as the 'weakest link' in the targeting process. Pressure to reduce human decision-making, in an *immediate* sense, from the planning and execution of targeting will therefore potentially increase on account of three push factors. First, a perceived advantage of granting full autonomy to AWS lies in their 'superior' ability to process a large quantity of information without facing human cognitive overload (Noone and Noone 2015, 33; Haas and Fischer 2017, 295). Second, AWS have 'agency' on the battlefield, in advantageous contrast to remote-controlled weapons that are steered by a remotely positioned, decision-making human, making such systems more vulnerable to countermeasures on account of transmission speed and susceptible to interference (Sparrow 2016, 96; Horowitz, Kreps, and Fuhrmann 2016, 26). Third, while AWS require the investment of considerable financial resources from development to testing and training, they are projected to turn out cheaper than human soldiers or pilot-flown jets in the long run (Gubrud 2013; Crawford 2016).

This trajectory has already become obvious even when only comparing, for example, the cost of manufacturing a single system: the US-manufactured Sea Hunter (I and II), autonomous submarine-hunting vessels (that at the time of writing remained unarmed) cost US$2 million, compared with the US$1.6 billion that acquisitioning an Arleigh-Burke-class destroyer entails (Scharre 2018, 79). Four Sea Hunter prototypes are expected to be under the command of Surface Development Squadron 1 (SURFDEVRON), established in 2019, which 'will be dedicated to experimenting with new unmanned vessels, weapons, and other gear to propel the surface force forward' by 2021 (Eckstein 2020).

Rather than only considering potential push factors for the development and deployment of AWS, we should also note that AWS come with operational disadvantages that may hinder their spread. Generally, it bears reminding that technological trajectories are rarely, if ever, linear, but instead subject to constant change and contestation. As an organisation, albeit not a monolith, the military is structured around control, which makes the idea of ceding it to AWS highly contested. The history of developing AWS in the US military, for example, includes various programmes that were cancelled despite considerable investment (Gubrud 2013) and even led to one report identifying 'a cultural disinclination to turn attack decisions over to software algorithms' (Watts 2007, 283). Even in a context that is overly enthusiastic about robotics and autonomy, these arguments still remain current. Therefore, even though the US Department of Defense's Third Offset Strategy is centred around deep learning systems and human–machine combat teaming as strategic components of retaining the country's technological advantage, '[t]here is intense cultural resistance within the US military to handing over combat jobs to uninhabited systems' (Scharre 2018, 61).

We will now summarise the extent to which these arguments have shaped the ongoing international debate on LAWS at the United Nations.

LETHAL AUTONOMOUS WEAPONS SYSTEMS AT THE UN

Since 2014, the international community has been discussing LAWS under the framework of the CCW. Following earlier advocacy work by ICRAC, this process took off properly with the publication of *Losing Humanity: The Case Against Killer Robots*, a report co-authored by Human Rights Watch and the Harvard International Human Rights Clinic (Human Rights Watch 2012), which set in motion a broader public debate on a potential, preventive ban of 'killer robots'. In 2013, Human Rights Watch became one of the founding members of the Campaign to Stop Killer Robots, an international coalition of non-governmental organisations dedicated to promoting a legal ban on LAWS. The same year saw the publication and presentation of another influential report on 'lethal autonomous robotics', authored by Christof Heyns, then UN Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions to the UN Human Rights Council (Heyns 2013). Here, many states present expressed the desire to continue discussing LAWS in the context of the CCW, which then started its first informal debates in May 2014.

The CCW is something of an unlikely forum in which to discuss such an emerging, highly publicised topic. On the fortieth anniversary of its adoption, in 2021, the CCW was not only positively cast as an '(unintended) incubator for ideas' but also, more cynically, as a 'place where good ideas die' (Branka 2021, see also Carvin 2017). Entering into force in 1983, the CCW was initially composed of an umbrella document as well as three protocols on weapons with non-detectable fragments (Protocol I), landmines (Protocol II), and incendiary weapons (Protocol III) (Carvin 2017, 57). Two further protocols have since been added: a preventive prohibition of blinding laser weapons (Protocol IV, entered into force in 1998) and explosive remnants of war (Protocol V, entered into force in 2006). Deliberations at the UN-CCW firmly frame the issue of LAWS as a potential problem for international humanitarian law, with civil society actors, chiefly the Campaign to Stop Killer Robots, hoping to convince states parties to commence negotiating new binding international legislation.

In 2016, the CCW formalised its deliberations through the creation of a Group of Governmental Experts (GGE) on LAWS. As of the time of writing in January 2021, the GGE had met six times, typically for five days at a time: in November 2017, in April and August 2018, in March and August 2019, and in September 2020. As the GGE mandate was renewed for a further two years in 2019, its discussions will continue at least until 2021. The GGE only has a discussion mandate, but similar mechanisms have, in the past, led to regulation agreements. The CCW has been and continues to be the only international deliberative forum where LAWS are substantially and regularly discussed. Therefore, it has become the focal point of transnational debate as well as of norm-promotion and lobbying activities of civil society actors.

This significance of the CCW for international debate on LAWS can be visualised by how many states have participated between 2014 and 2020. Seventy-five [out of 125] high contracting parties contributed formally[5] to the nine meetings on LAWS held between 2014 and 2020. Figure 1.2 goes into further detail, presenting data on the numbers of states parties contributing per year, categorised into three groups: Global North (GN), Global South (GS),[6] and states that support a preventive ban on fully autonomous weapons (ban share).

The data summarised in figure 1.2 point to three interesting observations (Bode 2019, 361). First, the number of formal contributions by states increased over time and across both the Global North and the Global South until 2018. The higher numbers for 2017 indicate a growing interest in the newly created GGE. Numbers stabilised in the four GGE years from 2017 to 2020, and we see the same countries continuing to participate in debates each year. However, both 2019 and 2020 participation numbers were below 2018 levels. This could indicate a growing fatigue with the lack of progress on the issue at the CCW, but only a review of the numbers for future meetings will confirm whether this signifies a trend; as 2020 saw the beginning of the Covid-19 pandemic, these numbers are not representative.[7]

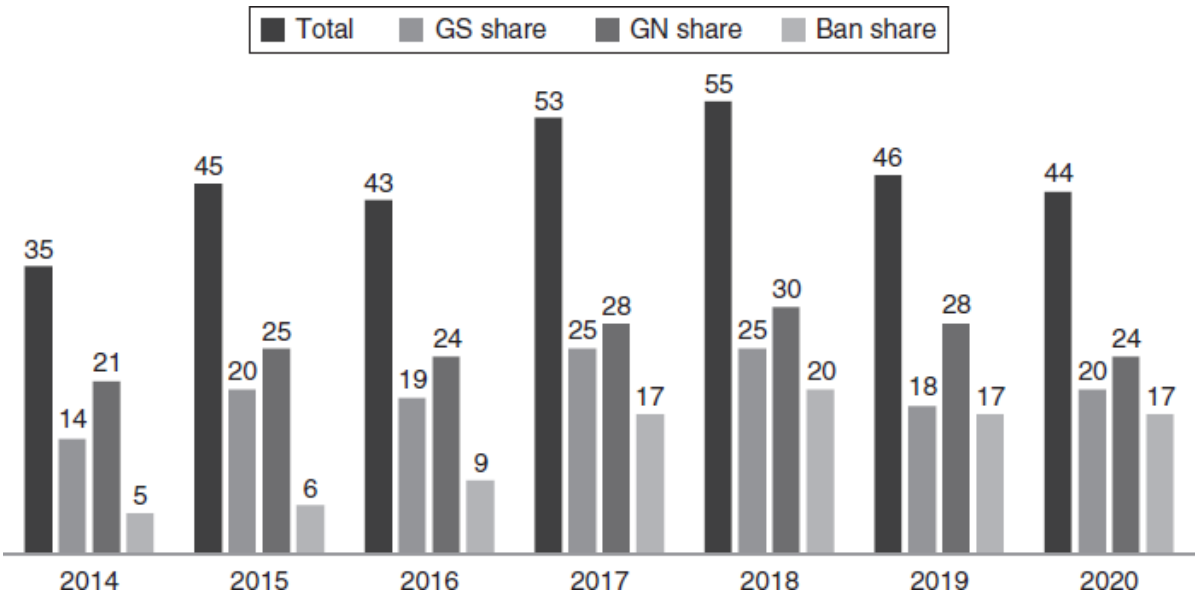

Figure 1.2. Formal contributions to debates on LAWS at the CCW (2014–20).

Second, the number of contributions by the Global South doubles when comparing 2014 and 2018. Third, the data also demonstrates the growing voice of ban supporters at the CCW: while only five supporters of the ban on LAWS contributed to the debate in 2014, this number grew to seventeen [out of thirty who supported a preventive ban[8]] in 2020.

Apart from Austria, who joined the 'ban group' in April 2018, this group of ban supporters is primarily composed of states from the Global South, and does not include any of the major actors developing such technologies – with the (possible) exception of China. At the GGE in April 2018, China became the first 'great power' calling 'to negotiate and conclude a succinct protocol to ban the use of fully autonomous weapons systems' (Kania 2018b). It has since repeated this commitment on several occasions. As others have argued, China 'wishes to ban the battlefield use of AWS, but not their development and production' (Haner and Garcia 2019, 335). However, doubts remain about the quality of its commitment.

Discussions at the CCW have been hampered by diverging viewpoints on autonomy and a resulting lack of conceptual consensus on LAWS. While some states parties have yet to share even a working definition, those put forward, for example, by the United States and the United Kingdom differ significantly. The United States defines LAWS as a 'weapon that, once activated, can select and engage targets without further intervention by a human operator' in US Department of Defense Directive 3000.09 (US Department of Defense 2012, 13). This language has become widely used in discourse on LAWS. But the Directive also includes a distinction between so-called semi-autonomous and autonomous weapons: semi-autonomous weapons can 'employ autonomy' for a full range of 'engagement-related functions' for 'individual targets or specific target groups that have been selected by a human operator' (US Department of Defense 2012, 14). The utility of this distinction is contested as target selection, under-defined in the Directive, becomes the only marker to distinguish semi-autonomous from autonomous weapons systems (Gubrud 2015), making it 'a distinction without difference' (Roff 2015a).

The United Kingdom differentiates between an 'automated' and an 'autonomous system': 'an automated or automatic system is one that, in response to inputs from one or more sensors, is programmed to logically follow a predefined set of rules in order to provide an outcome. Knowing the set of rules under which it is operating means that its output is predictable' (UK Ministry of Defence 2017, 72); in contrast, an autonomous system is defined as 'capable of understanding higher-level intent and direction. From this understanding and its perception of its environment, such a system is able to take appropriate action to bring about a desired state. 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' (UK Ministry of Defence 2017, 72).

In including the qualifiers 'higher-level intent and direction', this arguably defines away the challenges associated with short-term AWS. Further, the use of this specific phrase has political and legal implications: it sets a 'futuristic' and 'unrealisable' threshold for what qualifies as autonomy (Article 36 2018), is 'out of step' with how autonomy is defined by practically all other states (Noel Sharkey, quoted in House of Lords Select Committee on Artificial Intelligence 2018, 103), and makes it difficult to determine the United Kingdom's position on the use of less sophisticated AWS nearing development (H. Evans 2018; Article 36 2016, 2).

We summarise the substance of the debate at the CCW in the GGE years from 2017 to 2020 below, which also points to seven key changes and challenges. This summary also provides an account of negotiation dynamics and issues of substance raised at the GGE.

Key changes and challenges debated at the CCW from 2017 to 2020

First, in the light of the persistent definitional issues surrounding autonomy, various states parties as well as the three chairpersons of the GGE sought instead to move the debate towards exploring human–machine interaction. The concept of meaningful human control, also referred to *inter alia* as appropriate or effective control of LAWS, gained particular traction. As Brehm (2017, 8) notes 'what that involves, concretely, remains to be clarified' and still continues to be a matter of debate. There have been efforts by various states parties as well as civil society actors and independent experts to offer more concrete ways of defining meaningful human control, including expanding it to cover the entire life cycle of LAWS.

A prominent example of this includes a slide identifying so-called human touchpoints circulated by then chairperson Ambassador Amandeep Singh Gill of India. These touchpoints (also referred to as the sunset diagram due to the graphic it depicted) for the exertion of human control or supervision involved four stages: research and development; testing, evaluation and certification; deployment, training, command and control; and use and abort (Singh Gill 2018). This distributed approach to human control, to be distinguished from a previous focus on the use (and abort) phase only, has since gained traction at the GGE (e.g. Permanent Mission of the United States 2018b).

Other stakeholders in the debate, such as the International Committee of the Red Cross (ICRC), SIPRI, and the Campaign to Stop Killer Robots have also put forward detailed and practical operationalisations of meaningful human control (Boulanin *et al*. 2020; Campaign to Stop Killer Robots 2020a, see also chapter 5). Further, in 2019, the GGE included a reference to human–machine interaction in its so-called Guiding Principles (UN-CCW 2019) and discussions in 2020 continued to converge around this issue (GGE on LAWS 2020).

That said, the continued attractiveness of meaningful human control in deliberations may lie precisely in retaining its ambiguity, a recurring feature of (UN) diplomacy (Berridge and James 2003, 51). While GGE discussions on LAWS have not entered a negotiation stage, negotiations at the UN generally provide plentiful incentives for actors to leave core normative notions under-defined, vague or ambiguous, thereby creating the possibility for compromise (see, for example, Rayroux 2014).

Second, states parties' positions have become clearer, more substantial, and, incidentally, more polarised over time. At the first GGE in November 2017, voices of caution and critique dominated: no state party spoke explicitly in favour of developing LAWS, and many states parties followed a 'wait and see' approach. However, since the August 2018 meeting, some states parties (chiefly the United States but also Australia) have listed potential, and one must say supposed, benefits of LAWS in the area of precision targeting. Specifically, the United States, for example, argues that LAWS can strengthen compliance with IHL by effectuating the intent of commanders (Mission of the United States 2018c).

States parties to the CCW in 2020 can be more or less neatly split into three groups: the ban states (with Brazil, Chile, Costa Rica, Pakistan, and Austria being among the most vocal) calling consistently for an immediate transition to negotiating novel international law on LAWS in the form of a ban; the moderates (such as Germany, France, Switzerland), who are generally supportive of novel regulation but consider a political declaration or a strengthening of Article 36 weapons reviews rather than a legislative piece the most appropriate form at this stage of the debate; and the sceptics, including the United States, the United Kingdom, Australia, Israel, and Russia, who are united by their critical attitude on novel regulation rather than their positions of substance in the issue. As noted above, the group of sceptics includes states parties that argue expressly in favour of developing and using weapons systems with greater autonomy. But it also includes those, for example, Russia, who have cast doubts on the very existence of LAWS for years and since the start of the GGE process appear to be increasingly interested in slowing down discussions at the (often procedural) opportunities that present themselves. Other states parties, such as the United Kingdom, have argued that existing international humanitarian law is sufficient to deal with the technological evolution represented by AWS (Brehm 2017, 10).

Generally, over the period 2017–19 the atmosphere at the GGE grew increasingly polarised, which arguably had an effect on the positions expressed by academic experts observing the meetings.[9] We can underline this increased polarisation with the fact that finding consensus for the GGE's final report (a document with no legal character in the positivist sense) has taken increasingly more time since 2019. On the final day of deliberations, interpreting services stop at 6 p.m. due to financial/budget constraints. The states parties then move into a different room, where the remainder of deliberations and discussions on the final report take place in English only. While the discussion time was only exceeded by around two hours in November 2017, these discussions dragged into the early hours of the morning in both 2018 and 2019.

In the September 2020 meeting, the GGE debate saw less polarisation, with more states parties occupying a kind of middle ground and showing apparent interest in furthering discussion around the substance, in particular the extent to which IHL is sufficient, as well as aspects of human–machine interaction (GGE on LAWS 2020). That said, due to the current GGE mandate running until the end of 2021, states parties did not have to discuss the text of a final report.

Third, as noted above, states parties talk about very different technologies in the context of LAWS, making what is talked about deeply political: whereas some exclude existing weapons systems from the debate on LAWS, others stress that past and present violent practices involving mines, torpedoes, sentry guns, air defence systems, armed drones, and other technologies with autonomous features offer important insights into the changing modes and locales of human agency in the use of force and should be part of the debate (Brehm 2017, 13). A split has therefore emerged between those states parties that base their arguments in opposition to LAWS on these technological precedents and others who seek to clarify that the present discussions should only deal with 'fully autonomous' systems rather than so-called legacy systems or semi-autonomous system that they portray as acceptable (see also chapter 5).

Fourth, participants (and observers) of the GGE debate draw on fundamentally different images of technological futures, referred to as 'socio-technical imaginaries' in the academic literature (Jasanoff 2015). A first group embraces the development of LAWS, arguing that they represent a 'technological fix' for problems associated with current modes of warfare and that 'increasing autonomy in weapon systems enables conducting war in ever more moral and legal ways' (Brehm 2017, 14). Here, the inevitable deployment of LAWS is presented as an eventual empirical *fait accompli* that policymakers should simply make the most of by adapting adequate ethical and safety standards, not least for using 'AI for good in war' (L. Lewis 2019b; see also L. Lewis 2019a). By contrast, a second group challenges the supposedly inevitable journey towards LAWS (N. Sharkey 2012), highlights the fact that policymakers around the world have a decided scope for agency and choice in the matter, and, instead, problematises the ever-decreasing role of the human being in contemporary warfare. As Brehm (2017, 71) argues succinctly: 'Increasing autonomy in weapon systems is neither automatic nor inevitable. Inevitability is purposefully constructed by human agents. It is an ethical question and a political act when human agents attribute agency to a technological device or system rather than to people. This returns responsibility to us as representatives of institutions that deploy technology, who are involved in its design, who use the equipment or, perhaps most significantly, who are subjected to its operation'. This is exactly the line of reasoning we support in writing our contribution to the debate.

Fifth, civil society participation has always been strong at the GGE, in particular via the Campaign to Stop Killer Robots. In fact, the very discussion of LAWS in this forum and the creation of more formalised discussions in the GGE is entirely due to the civil society's successful lobbying efforts (see Bahcecik 2019). As the numbers of civil society organisations joining the Campaign have grown, so, therefore, have the number of contributions delivered as part of the GGE discussions. In 2018, for example, fourteen organisational representatives and academics spoke in favour of the Campaign's overall critical goal of preventively banning LAWS.

However, from 2019 onwards, there has been a sense of a growing disillusionment among many of the civil society participants. Having argued that technological developments are outpacing diplomacy since the beginning, civil society representatives increasingly voice frustration with the slow rate of progress at the GGE when it comes to deciding on a way forward – which is not least due to its consensual decision-making nature. Since 2019, suggestions to change the forum have therefore abounded. By moving outside of the CCW, civil society representatives hope to negotiate an international treaty outside of the UN's platform and therefore not be bound by its institutional challenges. (In fact, major treaties of international humanitarian law, such as the Mine Ban Treaty (also called the Ottawa Treaty), the Cluster Munitions Convention, and the Nuclear Ban Treaty, have been successfully negotiated outside of UN auspices in the past three decades.) The downside to this turn of events would be that major developers of LAWS, such as the United States, China, and Russia, are not very likely to be part of this negotiation process, thereby raising doubts about the eventual effectiveness of the potential future treaty in curbing LAWS development. But there are also important normative benefits to having a piece of international law that clearly expresses a moral desire to ban LAWS. Further, the global proliferation of drone technology over the 2010s to a point where at least 102 countries have military drone programmes underlines the potential for a similar arms race in the field of weaponised AI (Gettinger 2020). Importantly, this process demonstrates that not only the major military powers are relevant in proliferation processes: the list of countries with military drone programmes includes many with lesser military capabilities. A treaty negotiation process outside of the CCW is still likely to include their participation. At the same time, many campaigners still see a benefit in at least keeping major developers as part of the continued discussion.

Sixth, GGE discussions have often been led in a circular fashion, with some prominent themes re-emerging year after year. These issues become distractions or sideshows and are frustrating from a discussion standpoint as they slow down the pace of potential progress. Aside from the discussions around autonomy, two further issues have kept popping up in this way: LAWS and dual-use technology; and public opposition versus potential acceptance of LAWS.

Sceptics of entering into a negotiation phase with regard to LAWS have pointed to the dual-use nature of many of the technologies that sustain them as being a major obstacle when it comes to regulating them, fearing important technological innovation could be stifled. However, given that most military technology has elements that could be put to significant civilian use, including, for example, chemical weapons or blinding laser weapons, ways around this problem clearly could be, and have been, found if states parties are willing to consider them.

Concerning the supposed eventual public acceptance of LAWS as time passes by, campaigners have outlined how delegating the kill decision to machines violates basic standards of human dignity and the public conscience. This line of reasoning finds legal grounding in the Martens Clause. Such arguments are further sustained by successive public opinion polls, demonstrating that a solid majority of the general public (including in China, Russia, and the United States) oppose the development and use of LAWS (Campaign to Stop Killer Robots 2019a).

At the same time, proponents of LAWS or sceptics of their legal regulation suggest that the general public will eventually become more prone to accept LAWS as familiarity with AI in their daily life increases. The problem with this argument lies in its deterministic reasoning: why should increasing familiarity immediately lead to increasing acceptance when the reverse outcome is just as logical? In fact, the growing rollout of AI-driven technologies such as facial recognition has led to a more sustained public backlash over time. Further, there is a significant difference between using AI in weapons systems and using AI for functions not related to the use of force. As one interviewee noted: 'just because people are increasingly using laser pointers for their [PowerPoint] presentations does not mean that they are more likely to accept blinding lasers'.[10]

Seventh, when the format of CCW discussions moved towards a greater degree of formalisation with the GGE, many participants in the debate had hoped for a focus on more substantial discussion among states parties. The loose discussions on LAWS from 2015 to 2017 had instead been dominated by expert presentations. However, the first two years of the GGE (2017 and 2018) still included a significant number of expert panels, including presentations on seemingly abstract and far-away uses of AI that did not signal close proximity to the major matters at hand. Further, calls continued to be heard for more expert input, even in 2020. While expert advice should and does continue to complement GGE discussions, for example, via policy reports and GGE side events (held before and after sessions or during lunch breaks), giving more time to lengthy expert panel debates during the precious GGE debating time is not likely to be helpful in pushing the process forward.

In summary, the years of discussing LAWS under the auspices of the CCW have led to a greater clarity of purpose and a significant deepening of the substance of discussions, especially around issues of compliance with IHL and human–machine interaction. At the same time, the debate appears to have stalled since 2019 in the sense that the three blocs of states – proponents of a ban, moderates, and sceptics – remain more or less tied to their positions and there appears to be little room for moving forward on specifying the key notions of the debate. In the meantime, in many countries, technological development and planning for weapons systems with autonomous features in their critical functions continues to progress.

LITERATURE ON AWS: QUESTIONS OF LAW AND ETHICS

Starting in the late 2000s, AWS have been the subject of a lively scholarly debate, even if few of these contributions speak explicitly to the discipline of IR (Bode and Huelss 2019).[11] Research has explored AWS chiefly along two, often interconnected lines: questions surrounding their (potential) legality, and ethical challenges attached to their usage. This section summarises this debate, and comments on the gaps in this literature.

Studies of legality examine the extent to which AWS can or could potentially comply with international humanitarian law and, to a lesser degree, international human rights law. Chiefly, these legal concerns revolve around whether machines *can* technologically 'make the required judgment calls, for example, about who to target and how much force to use' (Heyns 2016b, 351). Below we tackle the concerns of international humanitarian law (IHL), international human rights law (IHRL), and *jus ad bellum* law. Often, scholars in this field draw on one of two analogies to frame AWS: they either discuss *instrumentally* weapons that can or cannot be used lawfully, or AWS are endowed with agency by capturing 'combatants' that can or cannot 'act' in adherence with IHL or IHRL (Crootof 2018). It is useful to read the following section with this critical framing in mind.

International law and AWS

All states are required to review any new weapons system that they seek to introduce, to demonstrate that the weapon itself can, in principle, be used lawfully. This is primarily based on Article 36 of the 1977 Additional Protocol to the Geneva Conventions. While not all states are parties to the 1977 Additional Protocol (the United States is a notable exception), its stipulations are considered generally binding as part of customary international law. Article 36 states that '[i]n the study, development, acquisition or adoption of a new weapon, means or method of warfare,' states are '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' (ICRC 1977).

These reviews should rule out the employment of weapons that are inherently unlawful for either of two reasons: if they 'cannot be directed at a specific military objective … or are of a nature to strike military objectives and civilians or civilian objects without distinction' (ICRC 1977, Paragraph 51 (4)); or if they 'cause superfluous injury or unnecessary suffering' (ICRC 1977, Paragraph 35 (2)).

Apart from providing clarity on legal requirements, there is some dispute in the literature about the role of human decision-making in weapons law. Some commentators note that these requirements do not immediately make the use of autonomous weapons systems unlawful if such systems are able to comply with international humanitarian law: 'the mere fact that an autonomous weapons system rather than a human might be making the final targeting decision would not render the weapon indiscriminate by nature' (Thurnher 2013; see also Anderson and Waxman 2013, 11). Along this line, it matters not who complies with these legal principles, be it a human or a machine, but instead whether they are applied correctly.

On the other side of the debate, human decision-making is described as essential in order to adhere to current standards of IHL compliance, 'i.e. a human commander must make a specific legal determination such as with proportionality' (Talbot Jensen 2018). Determining this would clearly entail investigating the precise role of human decision-making in international law. But the pivotal role of human decision-making can also be assumed to be located in the spirit, if not (necessarily) in the letter, of international law, resting on a monotheistic approach that privileges humans (Noll 2019). As Heyns (2016a, 8) argues, 'it is an implicit assumption of international law and ethical codes that humans will be the ones taking the decision whether to use force, during law enforcement and in armed conflict. Since the use of force throughout history has been personal, there has never been a need to make this assumption explicit'.

Important *jus in bello* principles governing the use of force – so-called targeting principles – raise fundamental concerns about the implications of increasing autonomy in complex aspects of weapons systems. Distinction and proportionality are key principles here and we will consider these in turn.

Distinction, i.e. the compliance with and ability to distinguish between civilians and combatants as well as between civilian and military objects, is the most fundamental principle of IHL as enshrined in Protocol IV of the Geneva Conventions and as part of customary international law (ICRC 1977, Paragraph 57 (2) (iii)). Distinction therefore guarantees civilian protection in clearly prohibiting their deliberate attack, making distinguishing between civilian and combatant an essential protective assurance of IHL.

Ultimately, the ability of AWS to adhere to distinction depends on the extent to which distinguishing between civilians and combatants is something that their targeting algorithms are capable of or the extent to which we see this as even programmable. Some scholars emphasise that whether AWS will be able to meet the requirements of distinction also depends on the contexts of their envisaged usage, differentiating between less and more complex contexts. In less complex contexts such as battles between declared hostile forces or in remote areas (e.g. underwater, desert) AWS 'could satisfy this rule with a considerably low level ability to distinguish between civilians and combatants' (Thurnher 2013; see also Anderson and Waxman 2013, 11).

We can see how this 'envisioning of AWS operating in empty spaces far away' (Brehm 2017, 40) explicitly and implicitly figures in graphics employed by diplomatic missions in Geneva. To illustrate, a presentation to the GGE meeting in August 2018 delivered by the Swedish delegation included a depiction of an anti-armour warhead to target tanks with autonomous trigger function in an empty desert setting. In selecting what is visible, these images arguably seek to shape what appears 'appropriate' in terms of AWS practices by limiting the imagination to their deployment in militarily clean, but unrepresentative, scenarios.

However, in complex environments, these requirements are considerably more demanding and scholars question whether AWS will ever be able to meet them (Sparrow 2016; Thurnher 2013; Asaro 2009). To argue that targeting algorithms may gradually become 'good enough', in whatever terms that may be defined, remains highly speculative as current programmers 'are a long way off, even in basic conceptualizing, from creating systems sufficiently sophisticated to perform … in situations densely populated with civilians and civilian property' (Anderson and Waxman 2013, 13). This finding gains even more significance when we consider broader trends in warfare that have seen war fighting in urban landscapes, as illustrated by the ongoing conflict in Syria, emerge as an environment that is characteristic of modern warfare.

Determining whether a vehicle is used for combat purposes or not, whether an armed individual is a fighter or a civilian, or whether a group comprising individuals also comprises civilians are questions of contextual human judgement. As these examples show, the legal definition of who is a civilian, for example, is not written in straightforwardly programmable terms, nor do machines have the necessary situational awareness and ability to infer required to make this decision (N. Sharkey 2010, 379). As former UN Special Rapporteur on Extrajudicial Killings, Philip Alston, summarised, 'such decision-making requires the exercise of judgement, sometimes in rapidly changing circumstances and in a context which is not readily susceptible of categorization' (UN General Assembly 2010b, Paragraph 39). Further, and importantly, although the targeting software included in AWS exceeds human capacities in terms of processing large data sets and other quantitative tasks, they are disadvantaged compared with humans when it comes to *deliberative* reasoning, interpreting qualitative data, or accurately judging complex social situations and interactions (N. Sharkey 2016).

The principle of proportionality allows civilians to be killed if their death is not deliberate and/or is justified by a proportionate response invoking military necessity, that is 'the military effects outweigh the unintended effects on non-combatants' (Kaempf 2018, 36). Distinction and proportionality therefore find themselves in an uneasy legal balance.

Again, scholarly opinion about AWS and proportionality differs and we encounter questions surrounding the human element that are similar to those related to distinction. Some regard AWS as potentially more precise weapons, thus decreasing human suffering (Arkin 2010) as underlined by the benefits of 'precision-guided homing munitions such as torpedoes' (Horowitz and Scharre 2014). Crootof (2015, 1879) holds that current weapons systems with autonomous features, such as close-in weapons systems (CIWS), are being used in adherence to the proportionality requirement. Further, she points to how US practice already uses a 'collateral damage estimate methodology' that is somewhat akin to pre-programming proportionality estimates (Crootof 2015, 1877).

Others argue that any decision about the proportional use of force requires assessing and processing of contextual and complex data that might be based on contradictory signals if measured against a preprogrammed set of criteria–action sequences of autonomous 'decision-making'. Based on this, some scholars argue that proportionality is even more difficult to comply with than other IHL principles 'because of its highly contextual applicability' (Laufer 2017, 71). These tasks pose significant challenges for AWS: 'To comply with the principle, autonomous weapons systems would, at a minimum, need to be able to estimate the expected amount of collateral harm that may come to civilians from an attack. Additionally, if civilian casualties were likely to occur, the autonomous systems would need to be able to compare the amount of collateral harm against some predetermined military advantage value of the target' (Thurnher 2013).

The fact that assessments of distinction and proportionality require highly complex reflection processes as well as value judgments therefore pose fundamental challenges to deploying AWS lawfully (N. Sharkey 2010). These considerations also highlight that, no matter how important advances in general forms of AI may be for future weapons systems, we should focus on 'stupid' autonomous weapons now (Bode and Huelss 2017).

The question of accountability has likewise seen significant coverage in the legal literature on AWS (Hammond 2015; Crootof 2016; Liu 2016; Jain 2016). This is often linked to public demands of individual and political responsibility when force is used, particularly in cases that violate norms of international humanitarian law (J.I. Walsh 2015). Failing to distinguish between civilians and combatants or using excessive force outside of the proportional assessments of military necessity constitutes such a war crime and triggers criminal liability (Sparrow 2007, 66). The spectre of such disastrous consequences looms particularly large in the case of AWS because of 'their destructive capacity and their inherent unpredictability' (Crootof 2016, 1350; see also Liu 2016, 330–31).

What distinguishes fully autonomous weapons systems is their operation without meaningful human control. In other words, AWS 'possess discretionary autonomy over the use of force' (Liu 2016, 328). More specifically, therefore, the accountability problem hinges on the absence of intent: 'by definition, war crimes … must be committed by a person acting "wilfully", which is usually understood as acting intentionally or reckless[ly]' (Crootof 2016, 1350). There may be cases connected to using AWS where commanders or other military personnel can be held directly or indirectly accountable – either because they ordered a specific use for the purpose of committing war crimes or because they did not abort an AWS when its use was likely to lead to war crimes being committed (Schmitt and Thurnher 2013, 277). But who is responsible for war crimes committed by an autonomous system if they result from unanticipated consequences of the use of AWS rather than the intended action of a commander in charge of the operation? Crootof (2016, 1375) therefore holds that the use of AWS may undermine a foundational principle of international criminal law as 'absent such a wilful human action, no one can … be held criminally liable'.

It is clear that 'legal obligations are addressed to human beings' who are 'accountable for harm done or infringements of the law' (Brehm 2017, 21). The increasing autonomy of weapons systems therefore also raises the question of the extent to which different groups of individuals such as engineers, programmers, political decision makers, or military command and operating staff are accountable for decisions undertaken and mistakes committed by AWS (Hammond 2015; Sparrow 2007; J.I. Walsh 2015). In the case of fully autonomous weapons systems – those operating without meaningful human control – scholars therefore speak of an accountability vacuum because 'it is uncertain who will be held accountable' (Heyns 2016b, 373).

Liu (2016) takes these questions one step further by identifying a conceptual gap between causal responsibility and role responsibility. The accountability gap discussed so far speaks to who holds causal responsibility 'for the unlawful consequences that are caused by AWS' (Liu 2016, 338). But there is also the notion of a 'role responsibility', tied to whether an individual adequately performed their role functions and obligations: 'a programmer has discharged his/her role responsibility if he/she has taken sufficient care to ensure that his/her algorithms function according to the requirement' (Liu 2016, 337). It follows that a human can have fulfilled their role responsibility in relation to AWS, but the use of AWS can still result in unlawful outcomes. However, the human cannot be held responsible for these outcomes because they can ultimately only be held accountable for failures associated with the inadequate performance of their role responsibility – thereby resulting in an intractable responsibility gap (Liu 2016, 339). Finding a solution to this can have detrimental consequences. As Liu (2016, 327) cautions, 'proximate human beings may become scapegoats in the zealous quest to plug the responsibility gap'. As we show in chapter 5, conceptually, this is arguably already the case in the context of human operators of air defence systems. While failures typically arise from complex human–machine interaction – that is, from how humans and machines operate together – disasters in the operation of air defence systems with automated and autonomous features are often blamed on human error. As Elish (2019, 41) argues, we may see the emergence of a 'moral crumple zone to describe how responsibility for an action may be misattributed to a human actor who had limited control over the behaviour of an automated or autonomous system'.

A final, more fundamental problem lies within the permissiveness of IHL when it comes to the use of force. While IHL provides an important limit-setting structure, e.g. it is not permissible to target civilians, it is permissible to kill civilians in missions that are militarily *necessary* and *proportional*: 'considerations of humanity require that, within the parameters set by the specific provisions of IHL, no more death, injury, or destruction be caused than is actually necessary for the accomplishment of a legitimate military purpose in the prevailing circumstances' (Melzer 2009, 77).

It is comparatively easy to find justifications for using force in IHL (Kennedy 2006). This permissibility is also visible in the particular context of new weapons reviews: 'IHL has been tailored to assist states in obtaining their desired military outcomes while employing weapons compliant with international law' (Laufer 2017, 64). This highlights that legal and ethical/moral concerns are fundamentally separate, and some deeply unethical types of behaviour are simply non-issues in IHL (see Scharre 2018, 6).

Some of the issues identified as concerns in IHL are also picked up in examinations of IHRL in the context of AWS. This body of literature is decidedly smaller in scale, 'probably because many commentators and policy makers envision the use of AWS in the context of military combat, rather than policing, and because discussions within the CCW are limited to the use of weapons as [a] means of warfare' (Brehm 2017, 11).

But in light of assessing whether autonomous weapons systems can be used in compliance with international law, adhering to the standards of IHRL is even more demanding: 'International human rights law is *much more restrictive about the use of force* than IHL. … Although law enforcement officials are allowed to use force under certain circumstances, such powers are strictly limited. They have a positive duty to protect the population and their rights, and deadly force may only be used to serve that purpose' (Heyns 2016b, 353; our emphasis).

Scholars raise significant doubts about whether machines could ever be programmed in a way to make the value judgments and assessments necessary to comply with IHRL, such as determining whether the use of force is necessary and whether a person represents an imminent threat (Heyns 2016b, 364–5). As Heyns (2016b, 366) summarises succinctly, 'there is a considerable burden of proof for those who want to make this difficult case'.

Importantly, IHRL not only makes it necessary to critically evaluate the use of force on a case-by-case basis, but also articulates demands beyond the potential application of lethal force to the decision-making processes involved in targeting: 'The algorithmic construction of targets draws on practices that are already considered deeply problematic from a human rights perspective, including secret mass surveillance, large-scale interception of personal data and algorithm-based profiling. The use of AWS is likely to sustain and even promote such practices, threatening human dignity, the right to privacy, the right not to be discriminated against and not to subjected to cruel, inhuman or degrading treatment and the right to an effective remedy' (Brehm 2017, 69–70).

Finally, although the literature on the international legal considerations of AWS is chiefly concerned with *jus in bello*, there are also a few contributions dedicated to *jus ad bellum* questions, arguing that AWS increase the general recourse to violent instruments, affecting proportionality and making escalations more likely. Roff (2015b, 41), for example, argues that even using AWS in self-defence against an act of aggression cannot fulfil the requirements of the *ad bellum* proportionality principle, understood as weighing the benefits/just cause of waging war against its overall consequences. Among these, she highlights that using AWS 'will adversely affect the likelihood of peaceful settlement' as the use of a publicly perceived robotic army will trigger deep resentment and further animosity among the population targeted (Roff 2015b, 47), as we have already seen in the case of drones (Kahn 2002; Oudes and Zwijnenburg 2011). In addition, the initial use of AWS will trigger an AWS arms race as 'other countries may begin to justify or view as necessary their possession and use of AWS' (Roff 2015b, 50). Other authors also identify this risk of escalation inherent in the unhindered proliferation and unfolding arms races involving AWS as a key negative consequence that would further lower overall thresholds to using force (Altmann and Sauer 2017; Sparrow 2009; Altmann 2013).

Ethics and AWS

Ethical studies cover a range of challenges associated with the potential deployment of AWS, centring on the question of whether autonomous weapons systems can ever *legitimately* harm humans or end human life (Johnson and Axinn 2013; Leveringhaus 2016; N. Sharkey 2010; Sparrow 2016; Schwarz 2018). This complements the legal question – can machines use force in compliance with international law? – with the more fundamental question – should machines make use-of-force decisions (Heyns 2016b, 351)? That is, even if using AWS were legal, could it ever be ethical?

Many call for a comprehensive ban on AWS due to ethical issues and concerns for human dignity (N. Sharkey 2010; Heyns 2016a; Rosert and Sauer 2019; A. Sharkey 2019), while a few voices, most prominently Ronald Arkin, emphasise their potential to contribute to more humane warfare, as AWS are not governed by emotions such as fear, hate, or revenge (Arkin 2009, 2010). Essentialising AWS as benevolent instruments often goes hand-in-hand with highly futuristic scenarios distinct from near-term advances in AI. Such studies also tend to neglect that even supposedly virtuous autonomous systems are only as 'good' as the purpose or intentions they are designed to serve. This links the current debate on technological autonomy and human–machine relations to general themes with regard to the ethical consequences of AI.

Before examining these ethical questions in the broader context of AI, technology, and human–machine interaction (which highlights some general problems), we discuss the main points of the contributions to the ethical debate on AWS. In general, the problems discussed here fall into the domain of applied ethics (practical ethics), which is an established research area in international politics. Applied ethics are directly linked to formulating guidelines, rules, and regulations that can be used in practical settings. In that sense, applied ethics are intended to translate normative content into practically useable standards of action. There are certain parallels here to norms but, in our understanding, norms are *broader* sets of standards that go beyond what is thought to be morally 'right'.

As mentioned above, one of the central, recent points of contestation is how AWS relate to human dignity. While the legal viewpoints on AWS are predicated on IHL and the principles of distinction, proportionality, and military necessity as benchmarks to debate whether the use of AWS could potentially be legal, ethical perspectives generally explore dimensions above and beyond the legal framework and consider whether AWS should be used even if they met principles of IHL. In this regard, authors tend to emphasise the lack of human deliberation and judgement (and, as consequence, the lack of accountability) that would violate human dignity if AWS used force to harm humans. It should also be noted that human dignity straddles legal and ethical areas of concern because much of IHL, as well as IHRL, is grounded in this principle (Heyns 2016b, 367). Briefly, 'underlying the concept of dignity is a strong emphasis on the idea of the infinite or incommensurable value of each person' (Heyns 2016b, 369). In the words of Asaro (2012a, 708): 'as a matter of the preservation of human morality, dignity, justice, and law we cannot accept an automated system making the decision to take a human life'.

In a highly informative take on the ethical questions raised by AWS, Amanda Sharkey identifies three arguments as dominant in the debate (A. Sharkey 2019, 78): '(i) arguments based on technology and the current and likely near future abilities of AWS to conform to IHL (i.e. what they *can* do); (ii) deontological arguments based on the need for human judgement and meaningful human control of lethal and legal decisions, and on considerations of what AWS *should* do. These include arguments based on the concept of human dignity; (iii) consequentialist reasons about their effects on the likelihood of going to war. These reasons include political arguments about their effects on global security, and are not necessarily labelled as consequentialist'.

Sharkey provides a detailed review of the human dignity argument in the context of AWS. She notes 'that there is a lack of a clear consensus about what dignity is' (A. Sharkey 2019, 82), which is, as Sharkey argues, the main problem affecting how concepts of human dignity influence the debate. In the same vein, Schippers (2020, 318) points out that '[w]hile dignity appears to offer a fixed-point for philosophical conceptions of human nature and a benchmark against which to judge policy, its meaning, as indeed the effects of its usages, are fuzzy and remain contested, partly as a result of the wider disavowed history of the term and its diverse set of historical sources'.

Nevertheless, human dignity figures prominently in the ethical debate on AWS, as summarised by Sharkey. The first argument holds that AWS are unable to 'understand or respect the value of life' (A. Sharkey 2019, 82). This argument is widespread in the literature, for example, in contributions by Christof Heyns: 'death by algorithm means that people are treated simply as targets and not as complete and unique human beings, who may, be virtue of that status, meet a different fate' (Heyns 2016b, 370; see also Asaro 2012, 2019; Rosert and Sauer 2019).

The second argument is that a human consideration of law-informing decisions is essential and that 'lack of human deliberation would render any lethal decisions arbitrary and unaccountable' (A. Sharkey 2019, 83). The third argument is that dignity is linked to a human rights 'package' and that 'AWS could affect all of these: limiting freedom, reducing quality of life, and creating suffering' (A. Sharkey 2019, 83). These are clearly valid arguments, but Sharkey also advises against arguing in opposition to AWS solely on the basis of human dignity due its ambiguous meaning. She goes on to argue that meaningful human control is a useful concept for considering the status of technologies and for assessing the extent to which they might violate human dignity.

We will return to a closer consideration and discussion of meaningful human control in chapter 5. In our view, current iterations of meaningful human control do not resolve the central point of contention revealed by ethical views on AWS: to provide a precise definition of an adequate involvement of humans and the exact quality of human–machine interaction. The argument that a sufficient level of human control builds the threshold that makes AWS acceptable requires, of course, that this threshold can be clearly defined. But is this possible? Positions in this regard vary, but we argue that even present forms of compromised human control as they can be found, for example, in the operation of air defence systems are ethically problematic, while they still set standards of appropriate use of force. We will elaborate on this point in chapter 5 and illustrate it empirically.

Raising a different point, Schippers (2020, 320) argues 'that the metaphysically anchored, dignity-based critique of AWS disavows the ontological and ethical entanglement of humans with autonomous and intelligent systems'. She proposes 'to read AWS through the lens of relational ethics … where the – real, perceived or projected – ontological qualities of autonomous systems generate conceptions of ethics, ethical responsibility, and ethical agency, and where human subjects, vis-à-vis practices of the self and collective engagement with others, constitute themselves as ethical subjects' (Schippers 2020, 322).

This perspective is in line with our emphasis on considering human–machine interaction as spaces where meaningful human control can be compromised. We argue here that increasing the technological 'sophistication' of weapons systems, including the integration of autonomous features, has crucial consequences. This does not refer to a potential future scenario but is rather an existing problem. Following the work of N.K. Hayles, Schippers (2020, 321–2) characterises her perspective of relational ethics in the following way: 'ethical agency is not based on the free will of an autonomous (human) subject, but emerges from the interpretation of information: human-technic cognitive assemblages interact with as well as transform the terms and terrain where ethical agency is exercised'. These are important, albeit complex and challenging, insights that share our emphasis on *practices* to study how and what kind of meaning is produced.

Turning to the other end of the spectrum, literature criticising the emergence of AWS on ethical grounds typically focuses on those who would be affected by using force. However, arguments in favour of integrating more autonomy into weapons systems also posit that such systems would increase 'precision' and therefore improve adherence to IHL principles, such as distinction, e.g. by enabling a better differentiation between a hospital and a military target (see Galliott and Scholz 2018). But viewpoints on human dignity underline that even using AWS in line with IHL can still be *unethical* (see Asaro 2012; Schippers 2020). As Schwarz (2018a, 25) notes, the focus on AI as a technological fix and solution to complex moral questions 'is ethics as a mere technical problem'. In that regard, ethical arguments in favour of using AWS are criticised for disregarding more fundamental questions of human involvement, such as those outlined above. This is also a problem of applied ethics that overemphasises the importance of law for providing ethical guidelines and 'seeks to establish certain ethical outcomes through regulatory frames, laws and codes' (Schwarz 2018a, 158), which can arguably also make ethics more accessible as a programmable tool for AI applications.

Another argument in favour of AWS as providing ethically superior outcomes holds that the use of AWS would be ethically imperative if it contributed to protecting own combatants (see Strawser 2010). Armed drones are a prime example of a weapons system that has reduced the role of personnel on the battlefield. But, while this removes combatants from harm, the risk has been transferred to civilians, because using remote-controlled technology could mean that the use of force is less precise or actually takes place where it would not have done if this technology were not available (see ICRC 2014, 18). While we can only estimate the numbers of civilian deaths due to the lack of transparency surrounding drone warfare, these are significant: The Bureau of Investigative Journalism (2021) posits that since 2010, between 910 and 2,200 civilians were killed by drones, among them between 283 and 454 children. As Schwarz (2018b, 286) argues, 'risk transfer from combatants to civilians in warfare is something that has been clearly observed in the US drone war. The foundation for this logic rests on the assertion that military lives matter just as much, if indeed not more, than civilian lives in warfare'.

Shannon Vallor offers an alternative take on AWS via the perspective of virtue ethics. This differs from deontological (rule-based, such as IHL) and utilitarian (such as an increase in or prevention of human suffering) ethical arguments that arguably dominate the discourse about ethics and AWS. Virtue ethics describe 'a way of thinking about the good life as achievable through specific moral traits and capacities that humans can actively cultivate in themselves' (Vallor 2016, 10). This leads Vallor to raise 'important ethical questions about robots that only virtue ethics readily allows us to pose: How are advances in robotics shaping human habits, skills, and traits of character for the better, or for worse? How can the odds of our flourishing with robotic technologies be increased by collectively investing in the global cultivation of the technomoral virtues? How can these virtues help us improve upon the robotic designs and practices already emerging?' (Vallor 2016, 211).

Virtue ethics give an interesting perspective on the impact that AWS can have on users: 'ethicists are starting to ask not just how robotic systems can be responsibly used by humans as tools of war, but also how robots themselves will alter, cooperate, or compete with the agency of human soldiers' (Vallor 2016, 212). Here, the central question is (Vallor 2016, 214): '[h]ow might the development of autonomous lethal robots impact the ability of human soldiers and officers to live nobly, wisely, and well – to live lives that fulfill the aspirations to courage and selfless service that military personnel pledge?'. Vallor (2016, 217) argues that AWS ultimately also threaten the virtues of the military profession, which are predicated on the central norm of selfless service that manifests in 'martial virtues' such as courage, loyalty, or honor, while acting in the framework of IHL.

In summary, research on ethics and AWS provides important contributions to the overall question our book seeks to answer – what kind of normative change can AWS induce? This debate draws our attention to the fundamental, ethical consequences of deploying AWS for those at the receiving end as well as for those actors 'using' AWS. But it does not systematically address the extent to which new standards of the 'appropriate' use of force emerge in in this context. Rather, the central question for ethical consideration is how our understanding and standards of protecting and upholding human dignity, human life, or ethical virtues change with the increasing importance of autonomous technologies or AI-driven decision-making in the military.

Beyond the specific context of AWS, there is also a much broader ethical debate about the role of AI, algorithms, and machine learning for human decision-making. The challenge of 'algorithmic regulation' (Yeung and Lodge 2019) concerns both government by and government through algorithms, as well as the regulation of algorithms in the public sphere. Yeung (2019, 22), for instance, identifies three normative dimensions concerning algorithmic decision-making systems: processes, outputs, and predicting and personalising services for individuals. Ethical considerations and contestations of increasing machine autonomy in decision-making focus on a similar range of issues in terms of input, processes, and outputs of human–machine relations. It is important to take these three dimensions to be interrelated but also to consider whether there is a hierarchy in their normative value. For example, an unethical process could compromise an arguably ethical outcome. In other words, arguments about a possibly superior algorithmic decision-making in terms of accuracy, precision, or reliability would still be overridden by violations of 'human dignity' as a superior ethical category.

In our view, research on algorithmic decision-making motivated by ethical perspectives centres primarily on procedural-consequentialist 'how?' questions. Arguments brought forward here highlight the lack or opacity of accountability and responsibility, as well as how it is not possible to appeal algorithmic decisions. In this regard, research on the ethical dimension of non-military AI has increased significantly in recent years. For example, the development of autonomous driving solutions raises a set of considerations symptomatic of the problem of robotic decision-making. Loh and Loh (2017, 37), for instance, provide a detailed discussion of the problem of responsibility, highlighting that responsibility is a relational concept consisting of five related elements: who is responsible; what *x* is responsible for; to whom *x* is responsible; the addressee defining the existence of responsibility in context; the conditions under which *x* is responsible. They argue that, 'as communication skills can vary, to say that someone is more or less able to act in a specific situation, more or less autonomous, more or less reasonable, and so on, it follows that responsibility itself must be attributed gradually according to the aforementioned prerequisites … assigning responsibility is not a binary question of "all or nothing" but one of degree' (Loh and Loh 2017, 38).

This concept of distributed responsibility, which also implies a potential hierarchy of more or less responsible agents (human and non-human), underlines the difficulty of dealing with the complexity of human–machine interaction in with regard to ethics. Even if we accept the moral and operational superiority of humans in distributed responsibility, the ethical problems remain complex. Bhargava and Kim (2017, 6) outline 'the problem of moral uncertainty' as follows: 'how should autonomous vehicles be programmed to act when the person who has the authority to choose the ethics of the autonomous vehicle is under moral uncertainty?'. This perspective contests that a technological fix can easily solve ethical issues that stem in particular from the important argument that 'robots are not agents' (Talbot, Jenkins, and Purves 2017, 258) and that ethical agency hence always depends on humans, who are, however, not necessarily capable of providing an adequate solution.

The essence of the ethical problem of autonomous driving is predicated on the possibility that AI causes harm to humans. This also entails, for example, the question of whom to protect during crash scenarios if different humans are involved as drivers or pedestrians, and to what extent this can be a life and death decision in ways not dissimilar to the ethical challenges raised by AWS.

Millar (2017, 20–34) highlights that 'according to Lin (2014), ethics settings in autonomous vehicles could reasonably be interpreted, both legally and ethically, as targeting algorithms …. Because collision management ethics settings involve decisions that determine collision outcomes well in advance of the collision, they appear very different from snap decisions made by drivers in the moments leading up to the collision. Seen this way, an ethics setting could be interpreted as a premediated harm perpetrated by whoever set it, which could result in that person being held ethically and legally responsible for the particular outcome of the collision'.

The question of responsibility and accountability is exacerbated by the increasing sophistication of machine learning in terms of self-learning systems, which already play a role in surveillance and targeting processes. The systems in question are able to perform 'specific forms of task intelligence' and 'in many cases they not only compete with but handily *outperform* human agents' (Vallor and Bekey 2017, 340, emphasis in original).

Considerations of ethical questions raised by autonomous systems in both military and civilian contexts often show that applied ethics run the risk of becoming colonised by the logic of technological 'solutionism' (Morozov 2014). As Schwarz (2018a, 165) notes, applied ethics 'seeks to use principles external to the realm it deals with in order to solve internal problems. This turns ethics in to a matter of problem-solving, for which a certain level of expertise is required to correctly identify and apply relevant external principles for a correct solution'.

This is an important argument that highlights the relevance of a type of situational ethics or, in other words, of flexibility and interpretative capacity. The notion of a human–machine assemblage as promoted by Schippers (2020) and Schwarz (2018a) could point in the direction of such an understanding, in that ethics overall cannot be programmed because judgements require a form of complex and sophisticated reasoning not yet delivered by AI. In a sense, we also adopt this line of thought in our view on norms as flexible and emerging in practice in contrast to being fixed and decided *a priori*.

While we will not further expand on the ethics of AI, the issues outlined above point to important broader debates and considerations currently taking place in academia, which often address similar problems as scholarship on AWS. We will focus on the contribution of this specific body of research in the following.

GAPS: WHAT IS MISSING FROM THE DEBATE?

Contributions to the study of AWS are rich and growing, but they currently come with two important gaps.

First, debates on the legality of AWS and ethical challenges attached to their usage both operate under a fixed, often predefined understanding of what is right and appropriate. However, we arguably require an analytical perspective that accommodates the constitutive quality of AWS as *emerging* technologies. In the following chapters of the book, we therefore present viewpoints on the flexible constitutions of appropriateness that come along with considering how AWS work through practices.

As is the case with all developing weapons technologies, there is no specific regulation in international law regarding the use of AWS. Further, while remote-controlled weapons such as drones have been used extensively by states such as the United States, their usage also remains largely unregulated. States continue to operate drones in contested areas of international law: drone warfare arguably violates the principles of distinction and proportionality, but is not covered by specific legal regulations (Kaag and Kreps 2014). Currently, the deployment of AWS remains legally permissible, with the added uncertainty that their technological autonomy is unaccounted for.

The most problematic aspect of the legal perspective on AWS is due to their elusive character. Any type of regulation requires looking towards the future in defining technologies that are constantly evolving. In other words, the technical complexity, the dual-use character, and the variety of deployment scenarios make it challenging to formulate legal restrictions on technological developments: 'History works against preventive norm-making. States usually react, as technological developments usually outpace political agreements to tackle future potential problems' (Garcia 2016, 101).

Yet, *standards of appropriateness* regarding the usage of AWS are already emerging. The increasing recourse to drones in the United States' approach to the use of force has led to the emergence of practices in the sense of routine ways of 'how things are done' (see Warren and Bode 2015). These are far from meeting the requirements of customary international law: they do not indicate 'dense' (behavioural or verbal) state practice, nor do they exhibit *opinio juris*, 'a belief that such practice is required … or allowed as a matter of law' (ICRC 2010). Yet, these practices create a *status quo* that makes it more difficult to challenge the further proliferation and usage of drones. This means that the emergence of customary law is even more restrained, despite the existence of practical precedents. The use and spread of the 'unwilling or unable' formula by the United States as a 'test' to inform drone targeting decisions is a case in point (Bode 2017a).

Accordingly, discussing this issue in the context of (customary) international law does not provide an adequate framework because it does not allow researchers to capture emerging standards of appropriateness attached to AWS. In contrast to considering how norms prevent or regulate, this book therefore studies how norms may emerge *outside of* deliberative forums and the framework of international law making.

Second, the incremental development of weapons systems with automated and autonomous features has impeded deep public discourse on this issue. There has been media attention, and civil society organisations or campaigns such as Article 36, the Campaign to Stop Killer Robots, and the International Committee for Robot Arms Control (ICRAC) seek to raise the issue's profile. But the wider political discourse is not particularly concerned with the most fundamental questions and realities of AWS that are significant for the quality of future human–machine interaction. Technological advances are typically incremental and come with a low political profile. Hence, the important role public acceptance plays in debating and creating dominant understandings of appropriateness is lacking.

Further, media coverage on AWS taps into fictional representations of AI, typically envisioning smart, humanoid killing machines (such as the T-900 from the movie *Terminator*) when it would be more appropriate to lead with real-life images (figure 1.3).

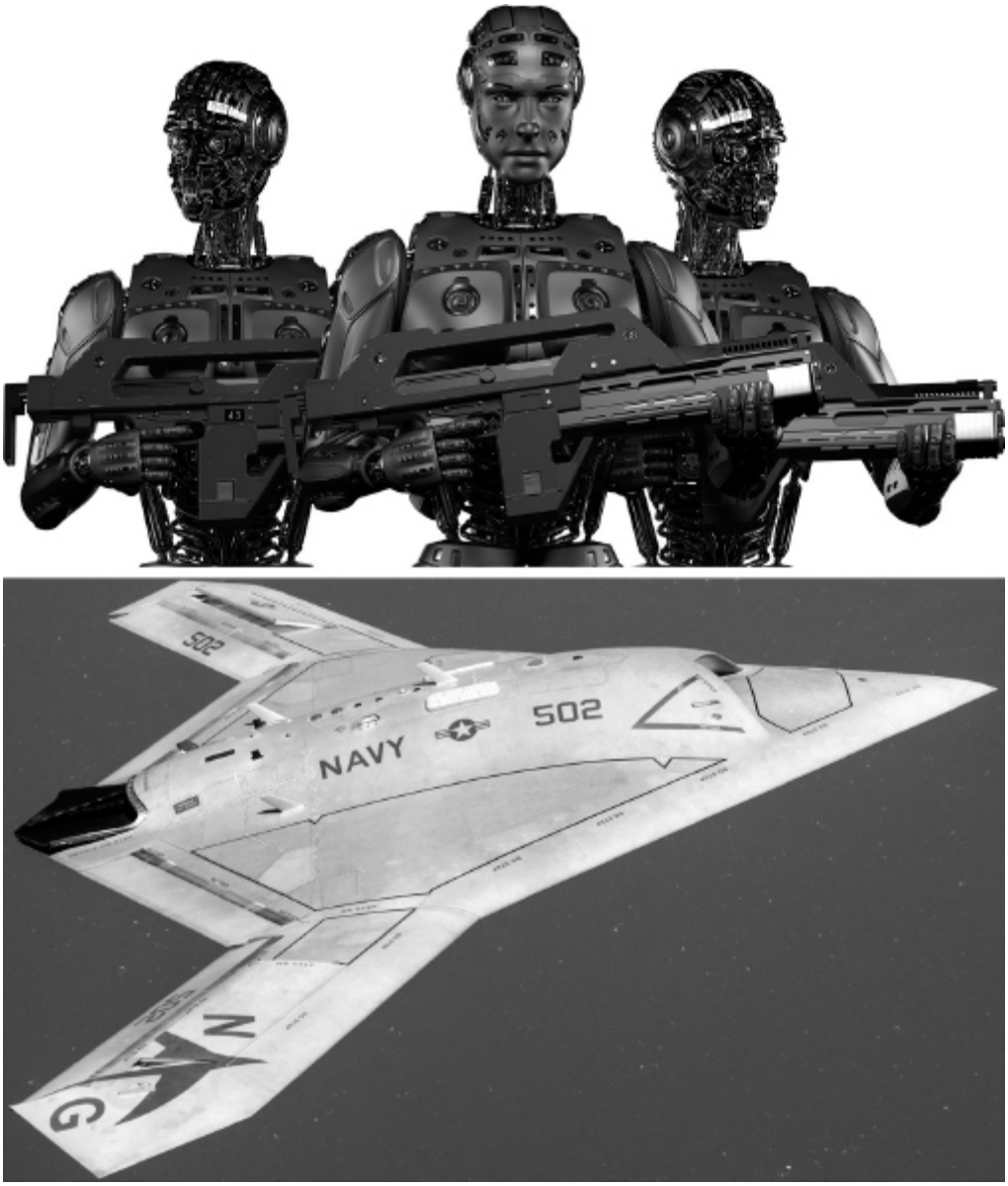

Figure 1.3 Humanoid killer robots (top) versus X-47B demonstrator with autonomous features (bottom).

Online editors have accompanied short pieces by the authors, for example, with sensationalist images of a robotic scorpion (Bode 2017b) or a humanoid robot hand about to press a red button (Bode and Huelss 2017). Clearly, such images are preloaded with meaning, and (fictional) narratives about AI and 'killer robots' shape the public imagination on AWS (Odell and McCarthy 2017). Such speculations about future technological trajectories, including those centring around the 'singularity', are rife in public debate but are undesirable, as they are often out of kilter with current capacities of weaponised AI (T. Walsh 2017b). This science fiction imagery serves to distance debate from what we should worry about in reality.

Building on the role of the public, the (lacking) discussion on developing AWS should also be seen in the context of overall lowering use-of-force standards when countering terrorism (see Bode 2016). Military efficiency and effectiveness as the main arguments for why drone warfare is appropriate have turned into a public legitimacy source (McLean 2014; M.W. Lewis 2013). The promise of 'surgical' strikes and the protection of US troops has turned drones into the most appropriate security instrument to counter terrorism abroad. AWS are 'considered especially suitable for casualty-averse risk-transfer war' (Sauer and Schörnig 2012, 375). This points to the important role AWS may play in democratic systems because they make the use of force appear more legitimate. In the case of drones, their broad acceptance across the military and political–public sphere marginalises their contested ethical and legal roles. A poll conducted in 2013 showed that 75 per cent of American voters approved drone strikes 'on people and other targets deemed a threat to the US' (Fairleigh Dickinson University 2013). The constitutive role that this security technology plays in international norms in terms of setting precedents and hence new standards of appropriate use-of-force is, however, completely out of sight.

To conclude, as the conventional frameworks of law and ethics have difficulties in accommodating flexibility and change for structural reasons, we have demonstrated that it is necessary to consider other types of norms that are not accounted for in research on AWS. A purely regulative perspective on AWS risks losing track of current developments and their future implications, in particular their possible role in shaping standards of appropriateness and our understanding of the 'new normal' in warfare.