Federal Foreign Office

Lethal Autonomous Weapons Systems

Technology, Definition, Ethics, Law & Security

Lethal Autonomous Weapons Systems

Technology, Definition, Ethics, Law & Security

The rapid evolution of new military technologies poses significant challenges. Autonomous weapons systems in particular are set to revolutionise the ways wars are fought. While completely autonomous weapons systems do not currently exist, already today certain critical functions in weapons systems are capable of operating autonomously. This trend towards gradually increasing autonomy in military systems in general and in weapons systems in particular will continue in the future. With the possibility of autonomous ground, air, surface, subsea, and space vehicles looming, it is likely to affect all domains of warfare.

Various significant strategic and operational advantages are associated with autonomous weapons systems. They are far more capable to adapt to and cope with the complexity, accelerated pace and data processing requirements of the modern battlefield than human soldiers. Factors that may cause stress, wrong decisions, or excess in human soldiers such as fear, anger, or hatred are absent in a robot. And they can perform the dull, dirty, and dangerous tasks and do so without exhaustion or any imminent risk to human life.

But the development towards autonomous weapons systems also carries significant risks. Concerns are varied and range from fears of a new arms race; qualms over unpredictable battlefield activities and resultant responsibility gaps; doubts as to these systems' ability to reliably abide by international humanitarian and human rights law; to ethical concerns over a devaluation of human life and dignity if life and death decisions are ceded to algorithms.

It is against the backdrop of these challenging questions that a third CCW informal meeting of experts on lethal autonomous weapons systems was held in Geneva in April 2016. Building on two preceding meetings in 2014 and 2015 that had already underscored the technical, ethical, legal and strategic questions raised by autonomous weapons systems, it was the objective of the third meeting "to discuss further the questions related to emerging technologies in the area of lethal autonomous weapons systems (LAWS), in the context of the objectives and purposes of the Convention of Certain Conventional Weapons". The present volume contains a collection of the expert opinions delivered at this occasion. They reflect a broad spectrum of views on autonomous weapons systems and, building on previous discussions, present a deeper probing in five focus areas:

First, a number of contributions in this volume address technological aspects of lethal autonomous weapons systems. Their objective is to map the spectrum of autonomous systems and to provide a clearer picture and understanding of the technological landscape. To this end, these contributions pursue the following questions: Where do we stand today? What are possible and realistic trends in the future? What challenges lie ahead? For example, how likely is it that we will see swarm applications of autonomous systems, deployments in increasingly diverse and complex scenarios or deep learning capabilities in autonomous systems?

Second, various contributions in this volume consider options for a (working) definition of autonomous weapons systems, examining concepts that have already been suggested with a view to further refinement and operationalization. Various military applications can be designed to operate autonomously and clearly not all of them would raise particular concerns or controversy. Drawing the line between those critical applications and areas where autonomy indeed raises concerns, and determining where full autonomy starts and ends, is a core challenge of the on-going debate about autonomous weapons systems.

Third, another focus area of the 2016 CCW informal experts meeting and the opinions collected in this volume is the legal dimension of autonomous weapons systems, with a particular emphasis on the laws of armed conflict, international human rights law and accountability issues. It is beyond any doubt that autonomous weapons systems that cannot comply with the laws of armed conflict must not be fielded. This means that weapons reviews and a clear understanding of whether and how autonomous systems can comply with international humanitarian law requirements are of crucial importance in relation to this new military technology. But even thorough weapons reviews cannot eliminate all risks. Wrongful conduct and accidents may still occur begging the question whether and how states and individuals could be held accountable when things go wrong.

Fourth, the deployment of autonomous weapons systems – even if such systems could be designed to fully comply with relevant legal requirements – raises fundamental ethical issues. Should machines ever be given the power to decide over life and death or to inflict severe injury, autonomously and without any direct human control or intervention? Would such a development towards mathematically calculated life and deaths decisions and the lack of empathy in such machines not undermine the principle of human dignity and challenge the principle of humanity as well as the dictates of the public conscience?

Fifth, autonomous weapons systems are widely considered as a transformational, sea-change technology that will have far reaching implications on all levels of strategic and military-operational decision making. Contributions in this volume therefore consider

a range of issues associated with the transformational effects autonomous weapons systems may have on international security relations. Among the issues addressed is the question whether the availability of such systems would lead to an increased likelihood of military interventions, the proliferation of autonomous technology including to non-state actors, the militarisation of the civilian technology sector, risk management and a potential loss of predictability, and the possible loss of public legitimacy due to accelerated decision making processes.

The debate and opinions about autonomous weapons systems continue to evolve. Discussions at CCW meetings in 2014, 2015 and 2016 have shown that there is a widely shared assumption that critical decisions may not be delegated to autonomous systems. Future discussions about these systems should therefore proceed from the understanding that where weapons systems are at issue, human control must be retained.

Robin Geiß

Table of Contents

Legal Challenges Posed by LAWS: Criminal Liability for Breaches of IHL by (the Use of) LAWS

Roberta Arnold*

Introduction

One of the major concerns raised by the (military) use of lethal autonomous weapons systems (LAWS) is the apparent uncertainty as to the attribution of potential harm caused by them in contravention to the laws of armed conflict (LOAC) or international humanitarian law (IHL), international human rights law (IHRL), or the *jus ad bellum*.

The present paper will focus on the specific issue of criminal responsibility for conduct amounting to a grave breach or to another serious violation of IHL, that is of the rules applicable in times of armed conflict that regulate the conduct of hostilities and the protection of specific categories of people (sick and wounded combatants, prisoners of war, and civilians), which are so serious as to entail individual criminal responsibility (also known as 'war crimes').[1]

Three key questions

Three key questions arise with regard to the issue of criminal accountability for breaches of LOAC/IHL committed by (the use of) LAWS:

1. Where does (international) criminal law fit within the international legal framework?
2. What are the LAWS-specific issues that arise under (international) criminal law?
3. Is there a need for new rules?

* Formal legal adviser to the Swiss Department of Defence, Laws of Armed Conflict Section, and former legal officer within the Federal Attorney General's Office, Centre of Competence for International Criminal Law Researcher, Stockholm International Peace Research Institute.

1 For an illustration of the evolution of this definition see R. Arnold, The ICC as a new instrument for repressing terrorism, New York 2004, pp. 67 et seq.

Where does (international) criminal law fit within the international legal framework?

The international legal framework is composed of different legal regimes: more than one may come into play to redress breaches of IHL committed by (the use of) LAWS. International criminal law (ICL) is just one of them.

a. ICL as one viable option for attribution

The first misbelief is that unless criminal responsibility can be attributed to someone for these breaches, no redress can be provided to the victims. Due to this, LAWS have often been condemned as 'killer robots' and viewed as a negative technological development, forgetting the fact that, under some circumstances, they can be more precise than other conventional means and actually spare lives. The question of attribution and redress for victims shall not be confused with the question of whether it is appropriate to 'delegate' the right to decide over someone else's life to a 'machine'. This is an ethical question, not a legal one.

So the first aspect to be aware of is that ICL may provide one among the various alternative and possibly complementary solutions available under international law to provide redress to the victims of breaches of IHL by (the use of) LAWS.[2]

In order to choose the most appropriate legal remedy, one needs to be aware of the various requirements and different objectives of the applicable legal regimes. For instance, the doctrine of state responsibility sets criteria for the attribution of conduct to *states*, and regulates the legal consequences of this attribution, including the duty to terminate unlawful conduct and to provide reparation for the resulting harm. Unlike *individual criminal responsibility,* state responsibility is not based on the notion of personal culpability, but on the attribution to the state of the (mis)conduct of its organs or agents.[3] And unlike state responsibility, criminal responsibility can only be attributed to an individual.

2 On this aspect see S. Sivakumaran, The Law of Non-International Armed Conflict, Oxford 2012; R. Arnold, Sivakumaran's Law of Non-International Armed Conflict: A Criminal Lawyer's Perspective, 48 Israel Law Review 253 (2015).

3 See R. Arnold, The Legal Implications of the Use of Systems with Autonomous Capabilities in Military Operations, in A.P. Williams and P.D. Scharre (ed.), Autonomous Systems: Issues for Defence Policymakers, The Hague 2005, pp. 83-97, http://www.act.nato.int/volume-2-autonomous-systems-a-transformation-in-warfare, p. 84.

Individuals may be held accountable under both criminal and civil law. The objectives, however, differ. Criminal responsibility shall only be attributed to those breaches of the law that are so serious as to justify a punishment and a criminal sanction. In times of armed conflict, this gravity is only given when the broken provision of IHL qualifies as a grave breach or another serious violations of the Geneva Conventions and their Additional Protocols, that is to a war crime.

b. The relationship between ICL and the other available legal regimes

The second aspect to be aware of is the relationship between ICL and the other legal regimes that exist under international law. The relevant legal regimes are:[4]

- jus ad bellum;
- jus in bello (i.e. LOAC/IHL), which applies to both international armed conflicts (IAC) and non-international armed conflicts (NIAC);
- international human rights law (IHRL);
- the law of weapons, which might be considered as a part of the LOAC, even though it partially addresses also arms control rules that already apply in peacetime;
- international criminal law (ICL).

For the purposes of the present discussion, the focus will be on the relationship between ICL and IHL.

IHL only applies in times of armed conflict, meaning that its breaches can only be committed during 'wartime'. 'Wartime' brings into play legal rules that take into account the exceptional situation and the specific needs of the affected states to counteract such situations, e.g. by resorting to the use of force and the intervention of their armed forces.

IHL is considered to be *lex specialis* and to take precedence over IHRL.[5]. Unlike breaches of IHL, breaches of IHRL do not qualify as war crimes, unless their content is restated in an IHL provision.

4 For an analysis see ibid., p. 86.

5 Even if the current trend is to increasingly expand the scope of application of IHRL to situations of armed conflict. On this specific aspect see the United Nations, International Legal Protection of Human Rights in Armed Conflict, 2011, http://www.ohchr.org/Documents/Publications/HR_in_armed_conflict.pdf, pp. 54-68.

ICL was created with the intention to attach criminal responsibility to the most serious violations of international law.[6] A new substantive (international) criminal law, composed of four main categories of crimes (war crimes, crimes against humanity, aggression, and genocide),[7] was established for that purpose.

The category of war crimes, which forms part of ICL, finds its roots in the grave breaches provisions contained in the four Geneva Conventions of 1949 (GCs) and their Additional Protocol I of 1979 (API).[8] As state parties agreed that the most serious breaches needed to be repressed with the mechanisms of criminal law, they introduced in the GCs the obligation for states to adopt the necessary provisions under their domestic laws. The four GCs of 1949 have been ratified universally, so that all State Parties, including those that did not ratify the Rome Statute of the International Criminal Court (ICC Statute), provide for criminal law provisions addressing grave breaches and other serious violations thereof.

With the development of ICL, the domestic rules implementing the grave breaches and other serious violations of IHL were later codified in international statutes and treaties such as the two Statutes for the International UN Tribunals for Yugoslavia and Rwanda (ICTY and ICTR) and, later, Article 8 ICC Statute. The latter is the result of a compromise, as the list of war crimes is not exhaustive (see e.g. the ban on chemical weapons, which was not included).

Therefore, it may be concluded that ICL is an offspring of IHL and its provisions regarding grave breaches: for their correct application, its provisions should be read in the light of IHL.

The war crime of unlawful killing, for instance, cannot be applied correctly unless one is fully acquainted with the general principles pursuant to IHL, namely distinction, proportionality, and military necessity. At the same time, it is also important to keep in mind the different setting, i.e. wartime: unlike IHRL, which primarily applies in peacetime, IHL does

6 As defined by Bassiouni: "International Criminal Law is a product of the convergence of two different legal disciplines which have emerged and developed along different paths to become complementary and co-extensive. They are: the criminal law aspects of international law and the international aspect of criminal law. The criminal aspects of international law consist of a body of international prescriptions containing penal characteristics, including criminalization of certain mechanisms. A study of the origins and development of the criminal aspects of international law reveals that it deals essentially with substantive international criminal law or international crimes"; see M.C. Basssiouni, Characteristics of International Criminal Law Conventions, in M.C. Bassiouni (ed.), International Criminal Law: Crimes, Vol. 1, New York 1986, pp. 59-80, p. 60.

7 The crime of terrorism, for instance, is not an international crime: it is a treaty crime. For more on this see R. Arnold (n 1), pp. 58 et seq.

8 Article 50 GCI, Article 51 GCII, Article 130 GCIII, Article 147 GCIV; also see Article 11(4), 85, and 86 API. See more references in R. Arnold (n 1), pp. 68 et seq.

not prohibit the targeting of a military objective by using lethal force, as long as its core principles are observed. This applies regardless of whether the military objective was targeted by a LAWS or a conventional weapon. Therefore, prior to attributing criminal responsibility for breaches of IHL by (the use of) LAWS, it is necessary to assess the context in which the alleged misconduct occurred, define the applicable legal regime(s), and if and only if the context qualified as an armed conflict under IHL, assess whether the elements of one of the war crimes established under ICL or domestic criminal law are met.

This is particularly important with regard to the breaches of IHL rules by (the use of) LAWS, since the fact that a system can use lethal force, as already mentioned, is not prohibited by IHL. Unless the use of lethal force, e.g. by LAWS, occurred within the context of an armed conflict, IHL and the war crimes provisions provided by ICL are not available as a legal remedy to the potential victims. This aspect is particularly relevant when discussing the legal implications of the use of LAWS within the context of law enforcement operations.[9]

As explained, when interpreting ICL provisions regarding war crimes, one shall take into consideration their history. Another important aspect to be taken into account for their interpretation is their objective. Notwithstanding their close nexus, ICL and IHL, in fact, pursue different objectives: IHL aims at regulating the conduct of hostilities (the so-called Hague Rules) and the protection of specific categories of persons, i.e. wounded and sick combatants, prisoners of war, and civilians (the so called Geneva Rules). IHL has a preventive and regulatory effect and aims at limiting the use of violence in times of armed conflict. It is not a criminal law treaty.

(International) criminal law, by contrast, aims at repressing and punishing a specific conduct. – in order to give teeth to IHL provisions concerning grave breaches and to ensure their compliance with general principles of criminal law such as legality, *nulla poena sine lege*, or *nullum crimen sine lege*.

When applying some principles and doctrines that can now be found in ICL, but which stem from IHL – such as the criminal law doctrine of command responsibility, which stems from the doctrine of responsible command under IHL – in order to interpret them, one must be aware of this principal distinction. For instance, under IHL and military law, the principle of responsible command aimed at ensuring a certain order and discipline among the armed forces, thereby facilitating the commander's duty to achieve his mission by retaining control over the troops and avoiding, among others, acts of insurrection or

9 This issue is beyond the scope of the present chapter.

violations of the applicable rules, including IHL rules. Failure to comply meant that the commander was going to be held accountable for his 'failure of duty' and not for the underlying acts committed by his subordinates.

The doctrine of command responsibility developed under ICL has gone a little further, and there is a tendency to assume that the commander failing to prevent or repress abuses of the law by his subordinates shall be held liable for such conduct. The present author, however, disagrees, and maintains that the doctrine should be interpreted in the light of its historical meaning. The consequence is that the commander shall only be liable for his failure of duty and that only the punishment should be proportional to the gravity of the underlying acts that he failed to prevent or to repress.

LAWS-specific issues raised under (international) criminal law for breaches of LOAC/IHL

Unlike the principles on state responsibility, (international) criminal law provides for the repression of war crimes committed by an individual.

An individual can be held criminally accountable only if he has met, by his conduct, the objective and subjective elements (*mens rea*) of the imputed crime. The judge will then appreciate the culpability and pronounce the appropriate sanction.

These basic principles of criminal law are posing an (apparent) problem in attributing criminal responsibility for breaches of IHL committed by (the use of) LAWS:

- LAWS do not qualify as 'individuals';
- LAWS are machines that cannot have *mens rea*;
- LAWS cannot be 'guilty' or 'culpable' of something;
- no matter what kind of sanction, this will have no impact on the LAWS.

For instance, as observed by the experts that attended the 2015 informal meeting on LAWS:

> *"Criminal offences are either caused intentionally or by negligence. When an autonomous system is intentionally misused or harm is caused through negligence, the human operator is criminally liable. When the* mens rea *(lack of intent, knowledge, recklessness or unjustified lack of awareness of the harm) of the user cannot be established there will be a responsibility gap."*[10]

The first question is whether this gap is specific to the use of LAWS; the second, if affirmative, is whether this gap may be filled by existing rules.

With regard to the first aspect, similar problems arise in relation to the criminal liability of enterprises, a concept that has been introduced in quite recent times. For example, Article 102 of the Swiss ordinary criminal code (CC) provides for the subsidiary liability of a corporation, when no individual can be identified as being responsible (for example as an organ of the corporation), due to deficiencies exiting in the structure and organisation of that particular corporation. In these cases, the corporation will be fined, so that the sanction will have an impact thereon. It would not make any sense, for instance, to pronounce a sanction providing for detention, as a corporation obviously cannot be detained.

Analogously, one may argue that the fact that LAWS are 'machines' that cannot have a personal *mens rea* and sense of guilt, and upon which a sanction like detention would not have the desired preventive or deterrent effect, does not mean that breaches of IHL committed by them (or their use) cannot lead to someone's criminal responsibility.

Research conducted within the framework of the MCDC project, led by NATO ACT in 2013-2014, on the challenges posed by the military use of autonomous systems to gain operational access,[11] showed that there is always a man 'in the loop'.

10 Meeting of the High Contracting Parties to the 2 June 2015 Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, UN Doc. CCW/MSP/2015/3, 2015 Session; Geneva, 12-13 November 2015, Item 8 of the provisional agenda: Lethal autonomous weapons systems, para. 46.

11 For more information, see the official public website at https://wss.apan.org/s/MCDCpub/default.aspx and the following literature on the project: R. Arnold (n 3), pp. 83-97; A. Kuptel and A. Williams, Policy Guidance: Autonomy in Defence Systems, 29 October 2014, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2524515.

This 'man in the loop' can either be found at the programming level or at the operating level: the programmer may be held accountable if he programmed the LAWS in a way to intentionally act in breach of IHL, for instance to target non-military objectives, the elimination of which was not required for military purposes.

At the operating level, the operator holding a degree of control over the LAWS may decide to operate it in an unlawful manner, thereby using the LAWS as a vehicle to commit a crime.

Alternatively, the 'man in the loop' may also be found at the deployment level: a commander might have told his subordinate to deploy the LAWS in an unlawful fashion, so that he will be the one directly accountable for a consequent violation of IHL.

Alternatively, if no direct perpetrator, that is person responsible for the conduct of the LAWS, can be found, the doctrine of command responsibility might be applied in order to attribute criminal responsibility for breaches of IHL by (the use of) LAWS. As already mentioned, this principle, which finds its roots in the military law concept of responsible command and the idea that the commander needs to retain command and control, as later codified in IHL, aims at attaching criminal responsibility to commanders for the misconduct of their subordinates, when these failed to fulfil their duty to maintain oversight and therefore to intervene in time when they acquired knowledge (or ought to have known, on the basis of the information available to them) that their subordinates were about to commit war crimes, either directly or indirectly via the use of LAWS. The criteria are clearly set out in Article 28 ICC Statute, which restates, even though in a slightly varied version, the principles contained in Article 86(1) API.[12]

Hence, for instance, if a commander had overlooked the fact that his subordinates were going to operate or deploy the LAWS in an improper manner, and the single individual who operated the LAWS unlawfully cannot be identified, one may resort to the doctrine of command responsibility in order to attribute criminal responsibility (if all respective criteria are met). Each case, to be sure, needs to be assessed specifically.

12 For an analysis of the historical development of this provision and its current meaning see R. Arnold and O. Triffterer, Commentary to Art. 28 ICC Statute, in O. Triffterer and K. Ambos, The Rome Statute of the International Criminal Court – A Commentary, 3rd ed., Munich 2016, pp. 1056-1106; R. Arnold and S. Wehrenberg, Die Strafbarkeit des Vorgesetzten nach Art. 264k StGB [The Criminal Responsibility of the Superior Under Article 264k of the Swiss Criminal Code], 2 Military Law and the Laws of War Review 19 (2013), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2532480.

Is there a need for new rules?

As observed by the independent experts that were convened during the 2015 informal meeting,

> "[t]his responsibility gap may not be a problem in the civilian domain if the social benefits of autonomous technologies (e.g. overall reduced accidents with autonomous cars), outweigh the social costs. However, this balance cannot be factored into the military domain and the assessment of the deployment of LAWS."[13]

In such cases, where neither a 'man in the loop' nor a responsible commander can be identified, the impossibility to attribute criminal responsibility to a person is not caused by the fact that the unlawful conduct was committed by a LAWS: the problem is rather the impossibility to collect evidence allowing for the identification of a single individual responsible for the misconduct. This problem is of a factual, evidentiary nature, not of a legal one. The same problem, however, may arise in relation to the use of conventional weapons. The LAWS might have passed all the tests and requirements set by Article 36 API and nevertheless be deployed or used in an unlawful manner by a person that cannot be identified.

This shows that the difficulties arising with regard to the attribution of responsibility for breaches of IHL committed by (the use of) LAWS are not LAWS-specific. The question that may arise subsequently is whether it is acceptable to entrust a LAWS with the targeting of an individual. This, however, is primarily an ethical, rather than a legal question.

In those cases, where no single individual can be identified, other applicable legal regimes and doctrines may provide for more viable solutions, such as the doctrine of state responsibility.

Conclusions

International law is composed of various legal regimes, *inter alia* IHL, IHRL, ICL, the law of weapons. Some of these are interrelated and may provide for alternative and complementary solutions to victims of breaches of IHL by (the use of) LAWS.

13 Meeting of the High Contracting Parties (n 10), para. 46.

Among these is ICL: this was developed on the basis of the obligation of states, as stated in IHL, to adopt the necessary domestic provisions to adequately address grave breaches and other serious violations of IHL (i.e. war crimes). The two regimes – IHL and ICL – are strongly interlinked: prosecution for war crimes can only occur if the breach was committed within the framework of an armed conflict. In recent years, especially in the aftermath of 9/11 and the beginnings of the 'war on terror', the borderline between the two scenarios – wartime and peacetime – has become blurred, and as a result the scope of application of IHL as well. Caution must therefore be taken in distinguishing potential breaches of IHL committed by (the use of) LAWS during an armed conflict or a law enforcement operation. In any case, LAWS are indeed subject to various regimes of international law, including IHL, ICL, and IHL adequately addresses new technologies such as LAWS. In this regard, there is no need for the emergence of new rules.

In order to attribute criminal responsibility for the commission of war crimes by (the use of) LAWS, the objective and subjective elements of the crime must be met. The difficulty to attribute criminal responsibility for breaches of IHL by (the use of) LAWS is often mistakenly associated with the fact that LAWS are machines that cannot have *mens rea*, and in regard to which the currently existing criminal sanctions do not have a preventive or repressive effect.

This assumption, however, oversees the fact that there will always be a 'man in the loop', either programming, planning, or deciding on the deployment of the LAWS, regardless of its degree of autonomy, or a responsible commander. As long as a nexus to this man in the loop can be established, and the analysis of the existence of the objective and subjective elements can be undertaken and determined, international criminal law will be a viable option for providing redress to the victims of a breach of IHL by LAWs.

If no single individual 'man in the loop' can be identified, the problem becomes of evidentiary nature. This difficulty, however, is not LAWS-specific.

Therefore:

- ICL is a viable option for providing redress to the victims of breaches of IHL by LAWS: LAWS are subject to ICL and IHL, and IHL adequately takes into account the technological developments;
- if no individual direct perpetrator/person responsible for the wrongful act in question can be identified, one might resort to the doctrine of command responsibility;
- in cases where not even the doctrine of command responsibility provides for a solution, legal regimes other than ICL might provide for better options, such as the doctrine of state responsibility.

Security, Unintentional Risk, and System Accidents in the Context of Autonomous Weapons Systems

John Borrie*

Introduction

When state policy practitioners in the context of the United Nations Convention on Certain Conventional Weapons (CCW) consider security and the increasing autonomisation of weapons systems, their natural preoccupation is with intentional actions and their consequences.[1] At the level of humanitarian and human rights law, for instance, a major concern is with accountability when machine systems are enabled to select and attack targets without a human decision in each individual attack.[2] At the level of non-proliferation and arms control, some experts are concerned that developments in autonomy will lead to arms races (that is, vertical proliferation) and horizontal proliferation – including to non-state armed groups.

Moreover, as Paul Scharre of the Center for a New American Security (CNAS) recently explained, autonomous weapons systems might be causes both of strategic stability and instability in crisis situations.[3] The use or threatened use of lethal autonomous systems might conceivably deter if it convinces an adversary that one's hands are tied in advance. It might

* Chief of Research at the United Nations Institute for Disarmament Research (UNIDIR), Geneva; associate fellow in international security at Chatham House. The views expressed in this chapter, which is based on a presentation to the CCW on 15 April 2016, are those of the author alone, and do not represent those of UNIDIR, the United Nations, or any other organisation.

1 Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), Submitted by the Chairperson, Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (CCW/MSP/2015/3), 2 June 2015.

2 For example, see C. Heyns, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Geneva, United Nations Human Rights Council, 2013, A/HRC/23/47, http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf; see also UNIDIR, Framing Discussions on the Weaponization of Increasingly Autonomous Technologies, No. 1, Geneva, 2014, p. 2.

3 P. Scharre, Center for a New American Security Presentation on 'Flash War': Autonomous Weapons and Strategic Stability, UNIDIR CCW lunchtime side-event on The Weaponization of Increasingly Autonomous Technologies: Understanding Different Types of Risks, Geneva, 11 April 2016. These and other presentations (including the author's) can be found at www.unidir.org.

allow a user to move up and down the escalation ladder more reliably and more visibly. But autonomy could also compound strategic fragility, for example due to inadvertent risk of various kinds.[4]

Inadvertent risk is examined in this chapter because it is a relatively under-explored subject in the context of security.[5] Thinking about inadvertent risk could usefully inform collective thinking overall about whether and to what extent lethal autonomous systems would pose unacceptable risk.

- Risk, simply stated, is the possibility of some bad event happening, something commonly quantified as the probability of an event multiplied by its consequences.[6]
- Unintentional risk denotes a subset of total risk in which machine systems that have autonomous targeting and attack functions fail to behave in ways intended – or necessarily even predicted – by their designers and operators.[7] There are many potential causes of inadvertent risk in autonomous weapons systems, and thus reason to believe such failures would be a question of 'when' and not 'if' such systems were to be developed and deployed.

4 Strategic stability is an important lens through which to see security, and the predominant one for at least some of the major contemporary military powers. However, it is important to note that strategic stability is, in itself, insufficient for security, even if it might be a necessary precondition in a multipolar world in which one major power cannot dominate the others.

5 Although there are exceptions. In particular, see P. Scharre, Autonomous Weapons and Operational Risk, Center for a New American Security, Washington D.C. 2016; see also J. Borrie, Safety Aspects of 'Meaningful Human Control': Catastrophic Accidents in Complex Systems, UNIDIR, New York, 16 October 2014; W. Wallach, A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control, New York 2015. Wallach's book is an important contribution that touches on these issues in the context of broader concerns about machine systems with autonomous targeting and attack functions.

6 See European Commission, Risk Assessment and Mapping Guidelines for Disaster Management, 2010, pp. 15-16.

7 UNIDIR, Framing Paper: Technology, Safety, and Unintentional Risk, UNIDIR Experts Meeting on the Weaponization of Increasingly Autonomous Technologies in Geneva on 7-8 April 2016, Geneva 2016 (available from the author on request).

Inadvertent risk

Let us consider some examples of potential failure in autonomous weapons systems that would contribute to inadvertent risk. Some have been discussed at the CCW meeting on lethal autonomous weapons systems, although there are others.[8] These possibilities can be categorised in various ways. In April 2016, for instance, a UNIDIR meeting of experts came up with several approaches. Due to space constraints, just one is displayed here (in Figure 1), with examples displayed in red. This diagram is illustrative and not exhaustive. Before the CCW's 2016 review conference, UNIDIR will publish an observation paper that explores these and related issues in greater depth.[9]

Figure 1 does not presume that a given system with autonomous targeting and attack functions would necessarily fail in these ways. But something the diagram displays is that potential causes of failure may reflect interactions between the machine system itself with the user or operator, as well as the overall context including the environment, the behaviour of adversaries and friendly forces, as well as the socio-technical system in which any technology emerges and is used. Indeed, a key point is that any of these many factors might interact with any other to compound the initial failure (you can think of these as connected by arrows). So when autonomous 'systems' are talked about, it is really not just the machine or the code itself that is solely relevant but the context into which it is embedded.

8 For instance, see presentations from the CCW Expert Meeting on LAWS from 11 to 15 April 2016, particularly those on mapping autonomy, human rights and ethical issues, and security issues, http://www.unog.ch/80256EE600585943/(httpPages)/37D51189AC4FB6E1C1257F4D004CAFB2?OpenDocument.

9 Issues around unintentional risk are to be explored further in a UNIDIR observation report on safety, unintentional risk and accidents in the weaponisation of increasingly autonomous technologies (2016, forthcoming) following a UNIDIR meeting of experts on these topics from 7-8 April 2016 in Geneva.

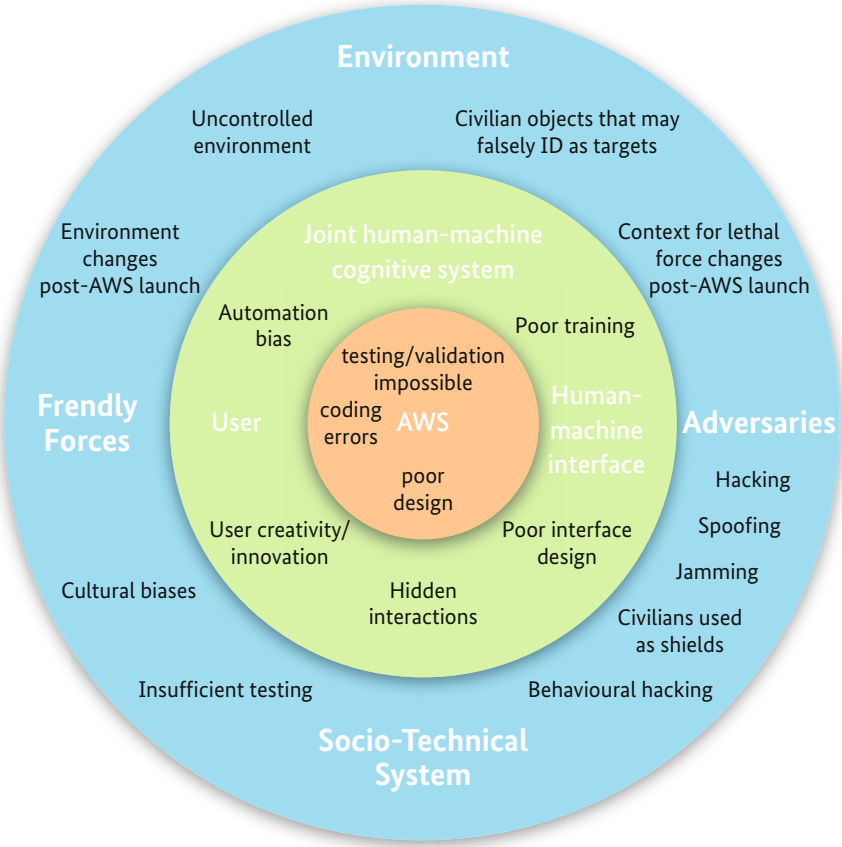


Fig. 1: Examples of cause of failure in AWS (courtesy of UNIDIR expert group on technology, safety, and unintentional risk, 8 April 2016)

In some cases, failures might have negligible consequences. In others, these sources of inadvertent risk could undermine the safe design and operation of autonomous weapons systems to the extent that they conceivably cause mass lethality accidents, and/or the creation of crises that undermine strategic stability.

System accidents

Moreover, so-called 'system accidents' are a special phenomenon to consider. These are important to discussions about the risks of autonomous weapons systems as they constitute a source of risk that cannot be entirely 'designed out' or quashed through rigorous operational standards.

The author has explored system accidents elsewhere (for instance at a UNIDIR side-event to the CCW on risk on 11 April 2016).[10] But here are just five points derived from the work of experts working in other areas, such as Charles Perrow[11] and Scott Sagan,[12] that are helpful in starting to think about the inadvertent risks of 'system accidents' with weapon systems that have autonomous targeting and attack functions:

1. *Accidents are inevitable in complex and tightly coupled systems.* In that vein, the experts that UNIDIR gathered from 7 to 8 April 2016 to discuss inadvertent risk concluded that autonomous weapons systems are likely to be both highly complex and very tightly coupled. In complex systems, there are a lot of common mode connections between components, which means it is not necessarily clear which of them has gone wrong when there is a failure. And there are many feedback loops. Tight coupling means that the consequences of a change or failure in one part of the system will rapidly propagate – perhaps quicker than human operators can respond.
2. *Safety is always one of a number of competing objectives (not least in war).*
3. *Redundancy is often a cause of accidents: it increases interactive complexity and opaqueness. It can also encourage risk-taking.* This is relevant to the discussion about autonomy, because the CCW discussions often focus on humans 'in the loop' as a form of ultimate redundancy, or machines as a form of redundancy for human decision-making. This may not reliably be the case.

10 J. Borrie, Unintentional Risks, UNIDIR CCW lunchtime side-event on The Weaponization of Increasingly Autonomous Technologies: Understanding Different Types of Risks, Geneva, 11 April 2016. See also J. Borrie, 'Safety aspects of "meaningful human control": Catastrophic accidents in complex systems', New York, UNIDIR, 16 October 2014, http://www.unidir.org/programmes/emerging-security-issues/the-weaponization-of-increasingly-autonomous-technologies-addressing-competing-narratives-phase-ii/weapons-technology-and-human-control.

11 C. Perrow, Normal Accidents: Living with High-Risk Technologies, New York 1984.

12 S.D. Sagan, The Limits of Safety: Organizations, Accidents, and Nuclear Weapons, Princeton 2013.

4. *The development of autonomous systems would not be value-free:* the preferences, biases, and assumptions of designers will shape machine systems even if these are intended to be highly predictable, reliable, and autonomous. It means that interactions with catastrophic risk potential may in practice be hidden from designers and operators.
5. *Hidden interactions might have particular risk potential in systems reliant on machine learning.* There are many different approaches within the field of 'machine learning'. Generally speaking, it involves 'training' a statistical model with some input data (normally in large volumes) so that the model produces useful output when presented with novel data. Some machine learning processes cannot be easily inspected or tested, but would be highly attractive in some military contexts in which, for instance, machines may have to function *incommunicado* from human decision makers for extended periods.[13] Problems such learning autonomous systems encounter, for instance in interpreting context, could lead to highly unexpected behaviour.[14] Such behaviour might include unintended mass lethality if the systems are armed and permitted to attack in at least some circumstances, or the pursuit of emergent yet inexplicable goals such as area denial to friendlies, as well as hostiles or neutrals. Either of these illustrative scenarios might have strategic consequences if they occurred in a crisis situation or conflict.[15]

Final thoughts

In sum, the propensity of complex, tightly coupled systems for 'system accidents' is relevant to security because these could undermine strategic stability in ways unintended by users and designers. These might be incidents of mass lethality, but failures in other ways (e.g. inadvertent area denial) could have strategic consequences as well. Such behaviour might not fall neatly into situations in which the law of armed conflict applies, despite such circumstances being the main focus of contemporary policy discussions in forums like the CCW.

13 For instance, in maritime environments. See UNIDIR, Testing the Waters: The weaponisation of increasingly autonomous technologies in the maritime environment, No. 4, Geneva 2015; see also D. Hambling, The Inescapable Net: Unmanned Systems in Anti-Submarine Warfare, Parliamentary Briefings on Trident Renewal, London 2016.

14 See, for instance, D.D. Woods, Chapter 10: Automation Surprises, in D.D. Woods and E. Hollnagel, Joint Cognitive Systems: Patterns in Cognitive Systems Engineering, Boca Raton 2006, pp. 113-142; more specifically on the 'context problem' see C. Yudkowsky, Artificial Intelligence as a Positive and Negative Factor in Global Risk, in N. Bostrom and M. Ćirković (ed.), Global Catastrophic Risks, Oxford 2008, pp. 308-345, p. 321.

15 Unintended area denial along shipping lanes in peacetime could also have strategic implications, as Collin Koh discusses in his chapter in this volume.

If Perrow's 'normal accident' theory holds[16], then system accidents are a cause of inadvertent risk and unpredictability that cannot be eradicated. Of course, this is the case with many contemporary hazardous technologies from nuclear power to nuclear weapon control systems[17] and manned spaceflight. We also know from these other domains, however, that while accidents may be rare, they nevertheless do occur. And when these complex, tightly coupled systems fail they tend to do so dramatically, if not catastrophically.[18] What arguably makes the problem of autonomous weapons systems different and even more challenging to consider is this: *failures with direct lethal consequences* could result from the decision-making of systems that are not human cognitive and moral agents *in addition* to resembling the kinds of complex, tightly coupled system that human society already struggles with in terms of safety.

Machine learning further suggests that the potential for hidden interactions in those systems would be greater – not lesser – because the processes by which they sense and interpret the world, and seek to achieve their goals within it cannot necessarily be predicted by human designers and operators. Common sense would therefore suggest that human decision-makers would actually want tighter control over such systems for that reason, which could run counter to the point of making them more autonomous in the first place. Indeed, both Perrow and Sagan have observed that centralisation (in this case human decision-making over targeting and attack) and decentralisation of control (autonomy) are often principles in tension in complex tightly-coupled systems.[19] However, until now there has not been the prospect of complex, tightly coupled hazardous technological systems beyond those governed by prescriptive rules that permit no deviations.

All of this also implies that there is a safety aspect to security, as well as to 'meaningful human control' or whatever the CCW eventually decides to call it. Among other things, the nature of systems with autonomous targeting and attack functions prompts the question of whether the requisite level of predictability, system transparency, training, and testing in order to allow the reassertion of human control before catastrophic consequences occur is realistically achievable – now or in the future. These safety issues should inform the CCW debate about security *and* meaningful human control.

16 See C. Perrow, Normal Accidents: Living with High-Risk Technologies, New York 1984.

17 See for instance J. Borrie. Risk, 'Normal Accidents', and Nuclear Weapons, ILPI-UNIDIR Vienna Papers, UNIDIR and the International Law and Policy Institute, Geneva and Oslo 2014.

18 For further discussion of this, and issues around human understanding of the problem of induction, see N.N. Taleb, The Black Swan: The Impact of the Highly Improbable, New York 2007.

19 See S.D. Sagan, The Limits of Safety: Organizations, Accidents, and Nuclear Weapons, Princeton 2013, p. 44. Sagan also offers a useful summary of Perrow's 'normal accident' theory from pp. 43-5.

Mapping the Development of Autonomy in the Military Sphere

Vincent Boulanin*

Introduction

Since 2013, the regulation of Lethal Autonomous Weapons Systems (LAWS) has been discussed under the framework of the 1980 United Nations Convention on Certain Conventional Weapons (CCW). The debate is still at an early stage, and it is clear that the implications of the trend toward increasing autonomy in weapon systems are not yet fully understood.[1] To help clarifying the discussion on LAWS at the CCW, the Stockholm International Peace Institute (SIPRI) in February 2016 launched a new research project intended to map the current development of autonomous capabilities[2] in military systems in general, and in weapons systems in particular. This article is based on early considerations and preliminary findings stemming from this work.[3]

Developments of machine autonomy in the military sphere today are, to a large extent, related to the use of unmanned systems, which has drastically increased over the last decade. This article therefore focuses on the progress of autonomy in this type of systems. The first section presents the military rationale for increasing their autonomous functioning. Section 2 maps out the current and foreseeable applications of autonomy. Section 3 reviews current hurdles to further adoption and use of autonomous capabilities in unmanned systems.

* Researcher, Stockholm International Peace Research Institute.

1 V. Boulanin, Mapping the debate on LAWS at the CCW: Taking Stock and Moving Forward, EU Non Proliferation Paper No. 49, 2016.

2 The concepts of automation and autonomy are used interchangeably in this article. Automation is considered as a spectrum, of which autonomy constitutes the higher end. Autonomy is defined here as the ability for a system to execute tasks to reach a goal in dynamic conditions with little or no input from a human operator. Reportedly, what distinguishes 'autonomous' functions from 'automatic functions' is the ability of the former to independently select its course of action from a number of alternatives to reach the goal. Unlike automatic systems, autonomous systems will not always yield the same output under the same input and conditions. See A. Williams, Defining Autonomy in Systems: Challenges and Solutions, in: A. Williams and P. Scharre (ed.), Autonomous Systems: Issues for Defence Policy Makers, Norfolk 2015, p. 33 (39).

3 The findings will be presented in December 2016.

Military rationale for increasing autonomy in unmanned systems: addressing the limitations of tele-operation

According to one author, autonomy characterizes the ability of a system to execute a task or several tasks without the intervention of a human operator, using behaviours resulting from the interaction of computer programming with the external environment.[4] In that regard, it is useful to approach autonomy as a capability attached to a system's function, rather than as a property of an object in itself.[5] Virtually any type of system, physical or computational, may be provided with autonomous features. Therefore, this article discusses autonomy in military systems rather than autonomous systems as such.[6]

Developments of autonomy in the military sphere today are to a large extent related to the use of unmanned systems, which has drastically increased over the last decade. Unmanned systems are now widely used for a large variety of missions, including Intelligence Surveillance and Reconnaissance (ISR), demining, logistics, and targeted strikes. The United States and Israel are the two countries that have made the greatest use of unmanned systems in military operations so far. Through its interventions in Afghanistan and Iraq, the United States has reportedly deployed more than 11,000 UAVs and 12,000 ground robots[7].

Nearly all unmanned systems that are in use today are tele-operated, in other words remotely controlled. Tele-operation is a model of human-machine interaction that ensures trust – since the human is constantly in the loop – but it has practical limitations. To begin with, it creates a dependence on high-speed bandwidth, which make the use of unmanned systems difficult or impossible in areas where communication is impaired due to environmental conditions or by adversaries.[8] Another limitation, which is fundamental for the military establishment, is that tele-operation is labour-intensive and thus expensive from a budgetary point of view. For instance, maintaining an UAV that is used for long loiter surveillance missions requires a rotating crew of pilots, sensor operators, mission controllers and a large number of data analysts.[9]

4 A. Williams (n 2).
5 United Nations Institute for Disarmament Research, Framing Discussions on the Weaponization of Increasingly Autonomous Technologies, UNIDIR Resources No. 1, Geneva 2014, p. 4.
6 For an overview of the different definitions of autonomous systems see I. Anthony and C. Holland, The Governance of Autonomous Weapons, SIPRI Yearbook 2014, Oxford 2014, pp. 423-431.
7 M.C. Horowitz, The Looming Robotics Gap, in Foreign Policy, May 2014, http://foreignpolicy.com/2014/05/05/the-looming-robotics-gap/.
8 United Nations Institute for Disarmament Research, The Weaponization of Increasingly Autonomous Technologies in the Maritime Environment: Testing the Waters, UNIDIR Resources No. 4, Geneva 2015.
9 P. Scharre, Robotics on the Battlefield Part I: Range, Persistence and Daring, Center for a New American Security, Washington 2014, p. 17.

Hence, for military planners, the future model for unmanned systems is their increasing autonomy. Allegedly, the goal is not necessarily to make them 'fully autonomous' and to take the 'human out of the unmanned system' altogether.[10] Rather, the objective is to find a model of human-machine teaming in which the human's and the machine's capabilities more effectively complement each other.[11] Autonomous functions would replace or support human operators in the execution of tasks for which their cognitive capabilities are not required or for which they are too limited. The human operators, on the other hand, would use their cognitive capabilities to make qualitative judgements that machines are not capable of. The relationship between the human and the machine would, consequently, be more dynamic. The control of the tasks, functions, sub-systems and even entire systems would pass back and forth between the human operator and the machine, depending on the requirements of the mission and changing circumstances.[12]

Current and foreseeable applications of autonomy in unmanned systems

Efforts to increase autonomy in existing military unmanned systems are occurring in several areas at the same time. The focus here is on three overarching areas of capability that are relevant for the discussions on autonomous weapons: self-mobility, collaborative autonomy and situational awareness.

Self-Mobility

Mobility is understood here as the capability of a system to govern and direct its motions in its environment. It is the area of capability where the interest for autonomy has been the strongest and consequently where technological progress has been the greatest. Discharging humans from the need to steer or pilot a system throughout the duration of a mission offers many advantages. To begin with, it reduces the cognitive burden on the human operator and thereby enables him/her to focus the attention on critical tasks. It also allows the military to increase their manpower efficiency as well as safety. Reportedly, the accident rate of UAVs during take-off and landing is lower when these tasks are operated autonomously by the machine rather than by a human pilot.

10 Department of Defense – Defense Science Board, Office of the Under Secretary of Defence Acquisition, Technology and Logistics, The Role of Autonomy in DoD Systems: Task Force Report, Washington 2012.
11 United States Air Force Office of the Chief Scientist, Autonomous Horizons, System Autonomy in the Air Force – A Path to the Future, Vol. 1: Human Machine Teaming, Washington 2015.
12 Ibid.

Progress in navigational autonomy has largely been conditioned by the complexity of the domain in which the system is intended to be used. It has been easier to achieve for UAVs and unmanned maritime vehicles (UMV) than for unmanned ground vehicles (UGVs), since the land environment is inherently more complex and cluttered.[13] For now, UGVs such as Israel's Guardium are currently only capable of navigational autonomy in environments that are well known, semi-structured and not contested.

Collaborative autonomy

Collaborative autonomy refers to the ability of a device or a system to communicate and collaborate with other devices or systems to reach a goal. Technological development in this area is still immature, but it is actively being researched. It is driven by the interest for multi-vehicle control and swarm robotics.

Multi-vehicle control gives a human operator the opportunity to control several platforms rather than a single platform at a time.[14] One emblematic research program for that type of capability is the CODE programme that was developed by the US Defence Advanced Research Projects Agency (DARPA). CODE stands for Collaborative Operations in Denied Environment. Its goal is to enable a group of UAS to work together under a single person's supervisory control: "The unmanned vehicles would continuously evaluate their own states and environments and present recommendations for coordinated UAS actions to a mission supervisor, who would approve or disapprove such team actions and direct any mission changes. Using collaborative autonomy, CODE-enabled unmanned aircraft would find targets and engage them as appropriate under established rules of engagement, leverage nearby CODE-equipped systems with minimal supervision, and adapt to dynamic situations such as attrition of friendly forces or the emergence of unanticipated threats'.[15]

Swarm robotics is an approach to multi-robot systems that consists of executing a task collaboratively by large numbers of simple and/or low-cost physical robots.[16] A number of experts foresee that developments in the field of swarm robotics will have a fundamental

13 M. Pomerleau, Work Shifts Down on Driverless Military Convoy, in Defense Systems, March 2016, https://defensesystems.com/articles/2016/03/31/bob-work-autonomous-ground-vehicles-unlikely-soon.aspx.

14 P. Scharre, Robotics on the Battlefield Part II: The Coming Swarm, Centre for a New American Security, Washington 2014, p. 37.

15 See Defence Advanced Research Projects Agency (DARPA), Collaborative Operations in Denied Environment (CODE), http://www.darpa.mil/program/collaborative-operations-in-denied-environment.

16 T. Tan and Z.-Y. Zheng, Research Advances in Swarm Robotics, in 9 Defence Technology 18 (2013).

impact on the future of warfare, as it would enable the application of force with greater mass, coordination, intelligence and speed.[17] Near-term applications of swarms include deploying of micro UAVs for ISR missions in cluttered environments,[18] confusing enemy forces in the air with small UAVs dispensed from larger inhabited or uninhabited aircrafts,[19] or overwhelming adversaries on sea with unmanned surface vehicles (USV).[20]

Situational awareness

Situational awareness can be defined as a system's ability to collect, process and analyse data on its environment to guide its decisions, those of the operator, or those of the military command. For the purpose of this article, the focus is on efforts to develop systems' situational awareness in relation to the targeting process, ISR, and electronic warfare.

Efforts to develop the capability of a system to identify, prioritize, select, and engage a target on its own are not new. Automated target recognition (ATR) is a technology that has been used for decades to aid human operators of manned and unmanned systems in finding and engaging enemies on the battlefield. ATR capabilities have proven to be particularly useful for the identification of targets beyond visual range of human operators and to track and engage fast moving targets, be it for offensive or defence purpose (e.g. missile and rocket defence). However, technological limitations in the domain of machine perception have been a major obstacle to significant progress in that area. Current systems can only recognise well-specified objects in uncluttered environments and under favourable weather conditions. They have little ability to learn and adapt to complex, novel and/or cluttered environments.[21] Further progress in that area is expected to come from future developments in sensing algorithms, but also in multi-agent collaboration, as it could facilitate the integration of sensors from different perspectives.

17 J. Arquilla and R. Ronfeldt, Swarming the Future of Conflicts, Santa Monica 2005; P. Scharre (n 14).

18 AFP, US Military's New Swarm of Mini Drones, in Defense News, 17 May 2015, http://www.defense-news.com/story/defense/international/americas/2015/05/17/cicadas-us-militarys-new-swarm-mini-drones/27494981/.

19 D. Lamothe, Veil of Secrecy Lifted on Pentagon Office Planning 'Avatar' Fighters and Drone Swarms, in The Washington Post, 8 March 2016, https://www.washingtonpost.com/news/checkpoint/wp/2016/03/08/inside-the-secretive-pentagon-office-planning-skyborg-fighters-and-drone-swarms/.

20 J. Golson, The Navy's Developing Little Autonomous Boats to Defend Its Ships, in Wired, 10 June 2014, http://www.wired.com/2014/10/navy-self-driving-swarmboats/.

21 J. Ratches, Review of Current Aided/Automatic Target Acquisition Technology for Military Target Acquisition Tasks, in 50 Optical Engineering 1 (2011).

The use of unmanned systems in long-loiter surveillance missions has also generated an interest for automating the data analysis process. As it stands, all the data that is captured by surveillance cameras integrated in unmanned systems has to be monitored and analysed by human analysts.[22] This setup is particularly labour-intensive and generates a need for a robust bandwidth. For military planners, the future of surveillance lies in the development of on-board computer vision algorithms that could identify situations of interest and cue them to human analysts. However, the progress of computer-assisted human recognition and understanding is slow. Existing computer vision algorithms still have very crude capabilities to recognise objects and to understand scenes in cluttered and unstructured environments.

Another area where the development of artificial intelligence is expected to leverage important benefits is electronic warfare.[23] While humans are still better than machines when it comes to image recognition, they have zero natural ability to perceive radio waves. They have to rely on machines to isolate unfamiliar radio-frequency transmissions. Many unmanned aircrafts do not have the capacity to detect new enemy signals on the fly, and when they do, they have to pass on the data to human analysts who can figure out a counter-signal to neutralise it. This process, as it is, can take weeks to months to years. Military planners are therefore interested in the development of autonomous jammers that could detect, analyse and counter new radio frequencies on the fly. Research in that area is still in its infancy but generates growing considerations. This year for instance, DARPA announced that autonomous radio wave identification and management would be the focus of a new research project named 'the Spectrum Collaboration Challenge'.[24]

22 P. Tucker, Robots Won't Be Taking These Military Jobs Anytime Soon, in Defence One, 22 June 2015, http://www.defenseone.com/technology/2015/06/robots-wont-be-taking-these-military-jobs-anytime-soon/116017/.

23 S. Freedberg, Jammers, not Terminators: DARPA and the Future of Robotics, in Breaking Defense, 2 May 2016, http://breakingdefense.com/2016/05/jammers-not-terminators-darpa-the-future-of-robotics/.

24 Defence Advanced Research Projects Agency, New DARPA Grand Challenge to Focus on Spectrum Collaboration, 23 March 2016, http://www.darpa.mil/news-events/2016-03-23.

Current obstacles to further advances of autonomy in the military sphere

Technological limitations

Certainly, the main obstacle to further development and use of autonomous capabilities in unmanned systems is the limitation of the technology itself. Major progress has been made at the hardware level. Sensors are increasingly precise, there have been significant advances in microprocessors which are increasingly powerful and small; parallel computing chips have enabled great strides in non-binary visual and auditory analysis; 3-D printing has made the production of actuators easier and cheaper. Telecommunications technologies can support a growing amount of information on the broadband. However, when it comes to developing autonomous systems, the real challenge is the performance of the software parts. With today's software technology, unmanned systems' ability to make sense of their environment and to adapt to changing situations is still very limited.[25] Major developments in the field of AI, notably in the areas of computational vision and machine learning, are needed to accomplish the military vision of what advanced autonomous systems should be able to do.[26] For now, autonomous systems are only used for narrow pre-programmed tasks in simple or predictable and unpopulated environments. The perspective of having soldiers and autonomous robots working in concert in offensive operations in complex and dynamic conditions still belongs to a distant future.

Cultural resistance

The limitation of today's technology is not the only obstacle to acceptance and adoption of autonomous systems by the military. There is also cultural resistance from within the armed forces. Military personnel are known for their reservations vis-à-vis certain new

25 A. Versprille, Army Still Determining the Best Use for Driverless Vehicles, in National Defence Magazine, June 2015, http://www.nationaldefensemagazine.org/archive/2015/june/pages/armystilldeterminingbestusefordriverlessvehicles.aspx.

26 M. Pomerleau, The Future of Autonomy Has a Strong Human Component, in Defense Systems, 23 Nov. 2015, https://defensesystems.com/articles/2015/11/23/future-autonomy-manned-unmanned-teaming.aspx.

technologies, particularly when their reliability has not been clearly demonstrated.[27] In the case of autonomous functions, there is a lack of trust that they will actually perform as intended, in all situations.[28]

Autonomy can also represent a threat to the very ethos of some military communities. Pilots, steersmen and logisticians may see the development of autonomy as a direct threat to their professions, as they feel that it incrementally makes their core skills and responsibilities obsolete.[29] It also has the potential to disrupt the operational paradigms they are used to. The possibility offered by distributed autonomy for multi-vehicle control could, for instance, lead to a shift from a model where a human operates one platform at a time to a model where a human operator would supervise and control several platforms in parallel. In that regard, a recent report found that the US military services showed different levels of preparedness to accept such an evolution. The US Air Force is allegedly showing particular resistance to the development of multi-aircraft control, while the Navy and to a lesser extent the Army are more enthusiastic about the operational possibilities that these technological developments could create.[30]

Inadequate defence acquisition process

Another important hurdle to the advancement of autonomous capabilities in the military sphere in the short term relates to how the defence acquisition process works in most arms-producing countries.

In general, ministries of defence's acquisition processes are ill-suited to the development of autonomous capabilities. They tend to focus on hardware, while critical capabilities provided by autonomy are embedded in system software.[31] Hardware and software require fundamentally different acquisition procedures. Typically, the acquisition process for hard-

27 M. Kaldor, The Baroque Arsenal, New York 1981; P. Dunne, Economics of Arms Production, in L. Kurtz (ed.), Encyclopedia of Violence, Peace and Conflicts, Oxford 2009; A. Toffler and H. Toffler, War and Anti-War: Making Sense of Today's Global Chaos, London 1995.

28 The Dwight Eisenhower School for National Security and Resource Strategy, Robotics and Autonomous Systems, Washington 2015.

29 Ibid.

30 P. Scharre (n 14), p. 37.

31 Department of Defense (n 10), p. 10.

ware stretches over decades, while the development of software requires rapid acquisition cycles and constant updates. The current models mean that software parts of new systems might already be obsolete by the time these systems enter into service.

Defence acquisitions processes also experience difficulty in quickly adopting new technologies from the commercial sector, which is clearly leading innovation in the fields of AI and robotics. Most defence procurement agencies are still looking for a reliable method to assess which commercial innovations have military potential, how to integrate them rapidly, and how to test and refine them.[32] Moreover, civilian companies have limited interest in developing solutions for the military sector.[33] Defence procurement processes are usually characterised by significant 'red tape' that puts off most non-defence companies. There are also concerns about long production cycles and intellectual property rights.[34] Moreover, there is reluctance by some actors in the commercial sector to see the military benefiting from innovations in the fields of robotics and AI. Some refuse to sign contracts with the military for ethical reasons.[35] Others are concerned about consumers' perception and the risk of bad publicity, especially since the revelations of the NSA leaks.[36]

32 On this topic see the speech of US Secretary of Defense Chuck Hagel at the Defense Innovation Days, Newport, Rhode Island, 3 September 2014. The US only last year created a Defense Innovation Unit, whose mission is to facilitate the import of commercial innovation in the defence sector.

33 G. Dyer, Robot Soldiers, in FT Magazine, 17 July 2015, http://www.ft.com/cms/s/2/45a2972a-2b2b-11e5-acfb-cbd2e1c81cca.html#slide0.

34 P. Tucker, As Pentagon Dwindles, Silicon Valley Sells Its Newest Tech Abroad, in Defense One, 22 April 2016, http://www.defenseone.com/technology/2016/04/pentagon-dawdles-silicon-valley-sells-its-newest-tech-abroad/127708/.

35 In 2015, the Future of Life Institute published an open letter against autonomous weapons systems, which was signed by over 8,600 people, including leading figures in the fields of AI and robotics. See 'Open Letter from AI and Robotics Researchers Against Autonomous Weapons, http://futureoflife.org/open-letter-autonomous-weapons/.

36 A. Mulrine, Pentagon Cybersecurity Strategy Comes With Olive Branch to Silicon Valley, in Christian Science Monitor, 23 April 2015, http://www.csmonitor.com/World/Passcode/2015/0423/Pentagon-cybersecurity-strategy-comes-with-olive-branch-to-Silicon-Valley.

Limitation of existing testing and evaluation methods

A final and fundamental hurdle to the adoption of greater autonomous technologies in the military sphere relates to testing and evaluation, and by extension legal reviews.[37] Ensuring that autonomous functions will perform as intended in a cluttered, dynamic and unstructured environment remains a fundamental challenge for the testing and evaluation community. A recent report by the Office of the US Air Force Chief Scientist noted in that regard that it is possible to develop systems having high levels of autonomy, but it is the lack of suitable validation and verification methods that prevents all but a relatively low level of autonomy from being certified for use.

Conclusion

To conclude, it is clear that unmanned systems will, in the future, increasingly rely on autonomous functions. Autonomous functioning generates opportunities for faster and more reliable task execution, it has the potential to discharge humans from dull, dirty or dangerous tasks, and can reduce the manpower burden that heavily impacts military budgets. Autonomy also enables greater reach, as it permits access to environments that are inaccessible to remote-control technologies.

The increase of autonomy in unmanned systems takes place in incremental changes, i.e. autonomous features are progressively added to each other. This does not necessarily mean that we see a linear development, towards full autonomy, with 'the man being taken out the unmanned' altogether ultimately – so to speak. For now, most technological developments are redefining human-machine interaction in a way that will permit humans to continue to exert control over the machine, but in a more flexible way. Human control will be exerted on some functions but not others, depending on the complexity of the mission, external operating environments and most importantly on legal and policy constraints. The question of what types of legal and policy constrains should govern the use of autonomous functions, in particular those related to the use of force, remains, however, to be debated both nationally and internationally.

37 V. Boulanin, Implementing Article 36 Weapons Reviews in the Light of Increasing Autonomy in Weapon Systems, SIPRI Insights on Peace and Security, No. 2015/1, November 2015, https://www.sipri.org/publications/2015/sipri-insights-peace-and-security/implementing-article-36-weapon-reviews-light-increasing-autonomy-weapon-systems.

Getting a Grasp of LAWS? What Quantitative Indicator-Based Approaches Could Bring to the Debate

Anja Dahlmann*

Introduction

In the context of the Convention on Certain Conventional Weapons (CCW), States Parties have been discussing the options for a regulation of lethal autonomous weapons systems (LAWS). Still, no definition of these systems has so far been found. In fact, it is even unclear whether the term LAWS comprises existing weapons such as Samsung's SGR-1, the South Korean military robot sentry, or Harpy, a weapon designed to loiter until it detects and destroys radar systems. These kinds of systems blur the lines between automated and autonomous weapons and are difficult to categorize, and therefore to regulate. The following chapter discusses several proposals for a working definition of LAWS, and highlights the possible benefits and flaws of indicator-based approaches.

The approaches discussed with regard to the international regulation of LAWS evolve quickly and become more complex and detailed every year, while many concepts are being debated in parallel or get mixed up. The approaches assessed below are not always independent of each other, but refer to or build on each other. The debate on the regulation of LAWS can be divided into two major categories of approaches. One is focussed on the machine's autonomy, or (as its counterpart) human supervision as whole. The other perspective considers different dimensions – and includes the element of risk, for example. While the more focused approaches are sleeker and easier to communicate, it is often unclear how to operationalise these concepts. That is where multi-dimensional and indicator-based approaches could help. Despite some flaws, they offer the opportunity for a structured

* Research assistant at Stiftung Wissenschaft und Politik – German Institute for International and Security Affairs (SWP), Berlin.

discussion on necessary (or critical) elements that should not excel a certain quality. The evaluation of these quantitative indicator-based approaches is based on experiences with the Multidimensional Autonomy Risk Assessment (MARA) Instrument.[1]

Qualitative Approaches

Over the last years, several approaches to define LAWS that were focused on the quality of autonomy or human supervision fuelled the debate on the regulation of LAWS.

First, the distinction between *automatic*, *automated* and *autonomous systems* tackles the definition on a rather basic level. Most authors refer to the complexity of the environment: while automatic and automated systems can only work in simple and structured environments, autonomous systems are able to act in complex situations.[2] For a working definition of autonomy, this might be too indistinct.

Second, instead of the complexity of decisions, the 'man-in-the-loop' concept focusses on the human contribution to the actions of a machine.[3] The status 'in the loop' requires full human command, while 'on the loop' means that the human controller can at least override the robot's actions, and 'out of the loop' equals robotic actions without (necessary) human input. It is a valuable model to visualise the different constellations of human-machine-interaction, but it does not include enough information on the actual human control, especially with regard to human-on-the-loop situations.

1 The first draft of this approach was published as a working paper in M Dickow et al., First Steps towards a Multidimensional Autonomy Risk Assessment (MARA) in Weapons Systems, 2015, http://www.swp-berlin.org/fileadmin/contents/products/arbeitspapiere/FG03_WP05_2015_MARA.pdf.
The MARA instrument is the result of a team effort by Christian Alwardt, Marcel Dickow, Niklas Schörnig, Frank Sauer, and the author, and builds upon an idea by Marcel Dickow (see e.g. M. Dickow, A Multidimensional Definition of Robotic Autonomy. Possibilities for Definitions and Regulation, April 2015, http://www.unog.ch/80256EDD006B8954/(httpAssets)/8FEA71BFEA5BBEE3C1257E28004149FD/$file/Dickow.pdf. The conclusions drawn in this text would not have been possible without the whole team's contributions. Nevertheless, the views expressed in this report are personal and the author's alone. The author is solely responsible for any errors in fact, analysis, or omission.
The author also wishes to thank the participants of SWP's workshop in April 2016 for their invaluable feedback on the MARA instrument.

2 See M.C. Horowitz and P. Scharre, An Introduction to Autonomy in Weapons Systems, February 2015, https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems, p. 6.

3 Most prominently Human Rights Watch, Losing Humanity: The Case against Killer Robots, November 2015, https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

Third, the efforts to define the term were followed by creating the concept of 'meaningful human control', introduced by the non-governmental organisation Article 36 in May 2014. The authors state that a situation in which the human is out of the loop always lacks meaningful human control, but that this lack can apply to an on-the-loop scenario as well. The main problem of this concept is its vagueness, which makes it politically appealing but at the same time difficult to operationalise.

In April 2016, Article 36 clarified its interpretation of the concept of meaningful human control by attributing it to several aspects of machine behaviour and the human-machine interaction. According to this updated definition, an autonomous system would have to behave predictable, reliable and transparent, and must allow for "sufficient confidence in the information that is guiding the human judgments being made, sufficient clarity of human action and potential for timely intervention and a sufficient framework of accountability."[4] Despite these very helpful additions to the concept, the scope of the term 'sufficient' remains unclear, as the authors themselves admit. Therefore, the problem of definition has not yet been fully solved.

In a similar line of argument, UNIDIR and others discuss the idea of 'critical functions':

> *"Not all autonomous functions are of equal concern: some might be uncontroversial while others raise significant legal, ethical, and strategic questions."*[5]

UNIDIR argues within the framework of meaningful human control, discarding technical definitions like the loop-concept or the distinction between automatic and autonomous. The overall assumption is that the concept of full autonomy is misleading with regard to the scope of the CCW negotiations.[6] Therefore, the debate should focus on critical functions like selecting and targeting, and should neglect functions like navigation and fuelling. However, the aforementioned problem remains: how to define and operationalise meaningful human control (in this case focused on a few selected functions)?

4 H. Roff and R. Moyes, Meaningful Human Control, Artificial Intelligence and Autonomous Weapons. Briefing paper for delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), Article 36, April 2016, http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf.

5 UNIDIR, Framing Discussions on the Weaponization on Increasingly Autonomous Technologies, 2014, p. 4.

6 See Chris Jenks at the CCW Informal Meeting of Experts on LAWS in April 2016, C. Jenks, The Confusion & Distraction of Full Autonomy, April 2016, http://www.unog.ch/80256EDD006B8954/(httpAssets)/7197832D3E3E935AC1257F9B004E2BD0/$file/Jenks+CCW+Remarks+Final.pdf; and in his contribution to this publication.

Multi-Dimensional Approaches

In order to acknowledge the complexity of the issue, other authors split up autonomy into different dimensions. For example, Horowitz and Scharre combine three (more or less) independent dimensions to analyse lethal autonomy. The human-machine command-and-control relationship describes the autonomy of a weapons system in "its relationship to a human controller",[7] while the complexity of the machine refers to the distinction between automated, automatic and autonomous. The type of decision being automated points in a similar direction as the concept of critical functions, but it can also include the complexity and risk of a certain task.

This multi-dimensional approach is not a comprehensive concept so far, but it adds clarity to the complex issue of autonomy.

Quantitative Indicator Based Approach

Apparently, it is hardly possible to pin down autonomy for political decisions or to put a threshold on sufficient human control. As it turns out, this vagueness is necessary to reach political consensus, but it might not be sufficient for an actual assessment. A quantitative indicator-based approach might help with that: It is a multi-dimensional approach, but evaluates interrelated aspects of a system and assigns certain values to them. The indicators describe functions of weapons systems that enable autonomous behaviour, but they could also include functions that might be especially problematic in connection with autonomy. This way, the assessment of the system could be tailored to a specific political problem that a regulation is supposed to address.

The idea of a quantitative indicator-based approach has been realised in the *Multidimensional Autonomy Risk Assessment* (MARA) in Weapons Systems. This instrument uses 14 indicators (called vectors) that are organised in five groups. Initially, it had two goals: it is supposed to allow for the comparison of different generations or versions of systems and it might help to define a threshold above which weapons systems are considered to be too autonomous and hence problematic (and therefore should not be used or even developed).

7 Scharre/Horowitz (n 2), p. 6.

The five groups of indicators used in the MARA instrument are physical characteristics, armament characteristics, human relevance, information processing or situational awareness, respectively, and exposition.[8] The instrument applies values to each vector on a scale from 1 to 10, and then combines these values to a formula. The outcome is supposed to model the specific risk of a weapons systems with regard to its level of autonomy:

> *"After scoring all vectors of a given weapons system, the MARA-formula generates the system's overall MARA-score that can be normalized to the percentage of the maximum value, the MARA%-score. By scoring and ranking a wide variety of weapons systems, MARA can generate a comprehensive, comparative overview against the background of which informed deliberations about a quantitative threshold can subsequently take place. This threshold, a specific MARA%-score, would have to be defined politically as the acceptable maximum of combined autonomy and military capabilities in weapons systems – or, in short, where to 'draw the line' for autonomy in weapons systems."*[9]

This approach tries to define which technical functions of an (autonomous) weapons system contribute to a specific risk, such as threats to international stability or international humanitarian law (IHL), and ethical problems under the assumption that only humans should decide on the killing of humans. Other challenges for the indicator-based approach in general could be the risk of proliferation or the violation of human rights. Before choosing the indicators, it is necessary to clearly define which risks the instrument is supposed to assess and help to regulate.

With regard to international stability, autonomous systems could be problematic due to their speed of reaction. They are much faster than humans and might not offer (sufficient) time for a human operator (or in this case supervisor) to stop their reactions. This could cause highly dangerous situations when it comes to unpredictable or erroneous behaviour in asymmetric conflicts, meaning the reaction of LAWS on manned or remote-controlled unmanned systems. It would be even more distressing and unpredictable with regard to the encounter of two or more LAWS. In addition to that, the threat perception of (potential) conflict parties could increase due to changing scenarios of weapons use following

8 For a full list of vectors and scales see Dickow et al. (n 1), p. 23-26.
9 Dickow et al. (n 1), p. 4.

the increased speed, higher range, and longer endurance in comparison to manned or remote-controlled weapons systems. In sum, many scenarios from conventional or nuclear deterrence also apply to autonomous systems, be it arms races or crisis instability.[10]

The MARA instrument tries to approach the issue of international stability through indicators for situational awareness and the possibility for human intervention. Situational awareness, in this case, includes the ability to interact with the environment as well as the mission tasking capabilities of a system. The possibility for human intervention is represented by the period of application. That would be the maximum time after which a human must check on the machine, which would be an opportunity to correct erroneous behaviour. In this sense, a solar-powered system with laser weapons would be more problematic than a battery-powered vehicle that only lasts a few hours. To assess the possible implications of a weapons system for international stability, the speed and range of the system are relevant as well. These vectors also interact with the type and range of armaments deployed on the autonomous platform with regard to mission (or threat) scenarios. Another indicator for the level of human control could be the time between deployment and use of weapons.

If an indicator-based approach is supposed to measure problems of international humanitarian law, it might be sufficient to focus on the situational awareness during the process of selecting and targeting of the weapons system. This would be based on the assumption that only a weapons system that is capable of interpreting and applying IHL should be left without human supervision, while every other system would need human control (or a very restricted area of application).[11] However, dealing with the errors or failures of the system, the possibility for human influence might be another factor to consider.

The indicators chosen for the MARA instrument to tackle these challenges have been mentioned above already: situational awareness and human intervention. To better assess the ability for compliance with IHL, it might however be necessary to operationalise these vectors in more detail.

10 See J. Altmann and F. Sauer, Speed Kills: Why we Need to Hit the Brakes on "Killer Robots", April 2016, http://duckofminerva.com/2016/04/speed-kills-why-we-need-to-hit-the-brakes-on-killer-robots.html.

11 It appears unlikely that LAWS will ever be capable of adhering to the principles of IHL by choice, see e.g. L Suchman, Situational Awareness and Adherence to the Principle of Distinction as a Necessary Condition for Lawful Autonomy, April 2016, http://www.unog.ch/80256EDD006B8954/(httpAssets)/F321892D-C9AE432CC1257F9A004A23FC/$file/2016_LAWS+MX+Presentations_Towardaworkingdefinition_Lucy+-Suchman+note.pdf.

The same goes for ethical considerations concerning the assumption that the decision to employ lethal force should only be taken by humans. Again, the aspect of human intervention comes into play. In addition to the aforementioned indicators, MARA scores the level of human involvement during the operational phase of a system through the vector 'actual human control', which basically represents an elaborated version of the human-in-the-loop concept.

Unmanned systems, currently mostly drones, are likely to proliferate, especially due to their dual-use character. The MARA instrument assesses the exposition (e.g. quality of encryption) of a system to evaluate the risk of manipulation or hijacking. Aside from that, the instrument does not include indicators regarding the risk of proliferation. It might be an interesting aspect to consider, but it seems as if such a scale would contradict the scales regarding international stability or international humanitarian law: the more complex and elaborated a weapons system is, the less likely it is to proliferate.[12] At the same time, the risk of destabilisation or violation of IHL might increase due to speed, range, or lack of human intervention in an autonomous and hence very complex system.

So far, the MARA instrument does not deal with human rights issues either. The threat to human rights by unmanned weapons systems has been prominently raised by Heyns, the UN Special Rapporteur on extrajudicial, summary or arbitrary executions, as well as by Amnesty International.[13] However, both focus on remote-controlled systems and their specific use for targeted killing, and it seems as if LAWS would not bring any new problems in that regard – at least if one assumes that the threat to human dignity is already covered by the ethical dimension.

As mentioned above, the MARA instrument might help to define a certain threshold to categorise unmanned weapons systems. The idea of a strong threshold that unambiguously divides 'good' from 'evil' weapons systems seems very appealing. It will however remain wishful thinking. First of all, such a threshold can be based on the mathematical model used for the respective approach, but ultimately, it has to be defined by a political decision. Secondly, a strict 'red line' could lead to the impression that systems below that threshold are unproblematic and can therefore easily be permitted. But since that must not necessarily hold true, it might be better to define a grey area or continuum in which the systems should be carefully discussed on a case-by-case basis. This shows that this

12 See K. Sayler (2015), A World of Proliferated Drones. A Technology Primer, 2015, https://www.cnas.org/publications/reports/a-world-of-proliferated-drones-a-technology-primer.

13 See C. Heyns, Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, UN GA A/HRC/23/47, April 2013; Amnesty International, Will I be next? US Drone Strikes in Pakistan, October 2013, https://www.amnesty.org/en/documents/ASA33/013/2013/en/.

type of approach will not directly lead to a regulation. Still, it can help to understand and evaluate specific risks of certain systems. The MARA instrument draws a cautious line at 50 percent of the overall MARA-score, but this has to be understood as a mere hint at particularly problematic systems.

A further possible use of the MARA instrument is that instead of creating a threshold, the values assigned to the assessed systems can also simply be used in order to compare them, for instance, by type, generation, or type of armament.

Flaws and Benefits of Indicator-Based Approaches

The MARA instrument received very helpful and detailed feedback from several scholars. The critique included very basic aspects such as the definition of risk, or the choice of indicators (including the impression that they do not always measure what they are supposed to), but also practical questions such as the inter-coder reliability. The latter aspect is of relevance in order to allow for a broad applicability among scholars or practitioners. A comprehensive codebook with clearly defined indicators, scales, and values could remedy that problem. Furthermore, discrepancies despite such a codebook could actually be helpful to stimulate an informed discussion about crucial aspects of the assessed systems.

These challenges would apply to other indicator-based approaches as well – in view of the assessment that there was the overall impression that the instrument could indeed be very useful to help better understanding the discussed weapons systems, and to see them in comparison with other systems.

As shown above, the level of human control (or at least the possibility to intervene) is a reoccurring theme and absolutely crucial for an assessment and regulation of LAWS. But since qualitative concepts such as 'meaningful human control' still lack a clear operationalisation, and full autonomy seems to be neither approachable nor helpful, quantitative indicator-based approaches could help to achieve a better understanding of autonomous functions and specific problems that arise from them. Therefore, the debate – and possible regulation – could benefit from a combination of qualitative approaches and quantitative indicators as guidelines for interpretation.

Cartographier l'autonomie des systèmes d'armes

Didier Danet*

Introduction

Les systèmes d'armes létaux autonomes (SALA) n'existent pas.

Si de tels systèmes devaient apparaître dans un futur encore indéterminé, nul n'est en mesure de dire quelles en seraient la nature exacte, les caractéristiques précis- es, les capacités ou les effets possibles de même que les dangers potentiels. Mais, il est cependant possible d'énoncer les conditions qu'un système d'armes devrait remplir pour être qualifié d'autonome. Ces conditions sont au nombre de trois et elles devraient être réunies cumulativement:

- un vecteur capable de se déplacer librement dans un environnement terrestre, aérien ou marin qui n'est pas entièrement connu d'avance;
- de réaliser le ciblage et le tir d'un effecteur létal (balle, missile, bombe);
- en autonomie totale de fonctionnement («human out of the loop»), c'est à dire sans la moindre intervention ou validation humaine («human in the loop») ou supervision humaine («human on the loop»).

Aucun système répondant à cette définition n'existe aujourd'hui. Mais, qu'en sera-t-il demain?

Les progrès spectaculaires enregistrés dans le domaine de l'intelligence artificielle et dans ceux connexes des télécommunications, du repérage dans l'espace, de la production d'énergie, des matériaux, ne risquent-ils pas d'aboutir inéluctablement et très prochainement à l'émergence de tels systèmes sur le champ de bataille? Ne faut-il pas s'en inquiéter et souscrire à la demande d'un certain nombre de Nations, d'organisations non gouvernementales[1] ou de personnalités qualifiées[2] d'en interdire le développement, la production, la commercialisation et l'usage?

* Pôle Action Globale et Forces Terrestres, Ecoles de Saint-Cyr Coëtquidan.

1 Human Rights Watch, Losing Humanity: The Case Against Killer Robots, Robots, November 2015, https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

2 S. Russell et al., Autonomous Weapons: An Open Letter from AI & Robotics Researchers, IJCAI, Buenos Aires 2015; J. Altmann et al., Armed Military Robots: Editorial, 15 Ethics and Information Technology 73 (2013).

Le débat se cristallise en particulier sur la question de l'autonomie que les pro- grès des sciences et des techniques pourraient conférer à la machine en l'affranchissant complètement de tout contrôle humain. Où en est cette autonomie? Comment et jusqu'où pourrait-elle évoluer dans l'avenir? Selon quel calendrier? Sous quelles conditions de coût? Comment la reconnaître et la mesurer? Nombre de ces questions sont encore sans réponse à l'heure actuelle. Un effort de cartographie peut donc apparaître comme un préalable à toute réflexion sur une hypothétique autonomie de décision et d'action.

Une telle démarche se heurte à des nombreuses difficultés. Dans le principe, il pourrait être tentant de s'inspirer des outils utilisés dans le cas des personnes physiques pour évaluer leur degré de dépendance, c'est à dire d'absence d'autonomie.[3] Mais, bien que des outils de ce type soient quotidiennement utilisés par les services sociaux français par exemple, leur transposition dans le domaine des systèmes d'armes soulève des difficultés majeures qu'il convient de ne pas mésestimer. Nous en évoquerons ici trois principales. La première est méthodologique et tient à la difficulté d'opérationnaliser les grilles d'évaluation: quelles questions poser? Quelles pondérations choisir? Quelles échelles de mesure adopter? Comment situer les progrès des sciences et es techniques sur ces échelles? La seconde est conceptuelle et résulte de l'imprécision des termes employés, la notion d'autonomie étant souvent confondue avec d'autres notions pourtant radicalement différentes, voire opposées, pour ce qui est de leurs effets sur le processus de décision militaire. Enfin, la dernière difficulté, et qui n'est pas la moindre, consiste dans le fait que le développement d'hypothétiques systèmes d'armes létaux autonomes n'aurait aucune utilité militaire si l'on veut bien considérer qu'elle contredirait les principes fondamentaux, politiques et stratégiques, qui gouvernent aujourd'hui l'action globale des forces armées.

Les quelques lignes qui suivent vont tenter, sinon d'apporter des réponses définitives, du moins de présenter les difficultés inhérentes à toute démarche visant à cartographier l'autonomie des systèmes d'armes. Il ne s'agit pas de dire que l'élaboration d'une carte de l'autonomie prospective des systèmes d'armes est par nature impossible. En revanche, il est important d'en mesurer les difficultés pour que le résultat obtenu ne soit pas totalement dénué de sens.

3 C. Martin, La dépendance des personnes âgées: quelles politiques en Europe?, Rennes 2015.

La cartographie de l'autonomie des machines: une méthodologie perfectible

L'évaluation du degré d'autonomie d'une personne physique est une question qui s'est posée de longue date, notamment dans le domaine de la protection des majeurs vulnérables ou dans celui de l'attribution des aides sociales aux personnes âgées. Pour ce qui est des majeurs qu'il s'agit de protéger en les plaçant sous un régime de tutelle ou de curatelle, la procédure judiciaire fait intervenir une expertise médicale individualisée.[4] La technique d'évaluation est donc celle d'un diagnostic circonstancié dont le degré de standardisation est relativement limité. Ce type de démarche pourrait s'appliquer à l'étude des machines mais la liberté laissée à l'évaluateur pourrait introduire un biais fort de subjectivité dans la démarche et les résultats seraient sans doute sensiblement différents selon les experts retenus pour procéder à l'évaluation.

En revanche, l'évaluation de l'autonomie réalisée dans le cadre d'une procédure d'attribution d'aides sociales fait l'objet d'une analyse objectivée, réalisée à travers la mise en œuvre d'une grille de cotation qui pourrait assez facilement se transposer au cas de l'autonomie d'une machine, du moins dans son principe.

Cette grille dite «AGGIR» (Autonomie Gérontologique Groupe Iso-Ressources) permet d'évaluer l'inverse du degré d'autonomie d'une personne, c'est à dire son degré de dépendance, afin de déterminer le type et le niveau d'aide qui peut lui être apportée.[5] Les niveaux de dépendance sont classés en six groupes iso-ressources qui vont de l'autonomie complète à la dépendance complète. La grille comporte dix activités corporelles et mentales dites «discriminantes», qui vont permettre de définir le groupe ico-ressources auquel rattacher la personne : communiquer verbalement, se repérer dans l'espace et le temps, faire sa toilette, se déplacer à l'intérieur du lieu de vie, utiliser un moyen de communication comme le téléphone... Ces deux activités font elle-même l'objet de mesure à travers un nombre plus ou moins grand de questions concrètes portant sur les actions du quotidien des personnes âgées. La grille comporte également sept catégories dites «illustratives» consacrées aux activités domestiques et sociales (préparer ses repas, respecter l'ordonnance du médecin, utiliser volontairement un moyen de transport collectif...) et qui, sans influer sur le niveau de

4 T. Fossier et al., Les tutelles: accompagnement et protection juridique des majeurs, Issy-les-Moulineaux 2015.

5 V. Coutton, Évaluer la dépendance à l'aide de groupes iso-ressources (GIR): une tentative en France avec la grille AGGIR, Gérontologie er société, 4/2001, pp. 111-129 ; S. Lafont et al., Relation entre performances cognitives globales et dépendance évaluée par la grille AGGIR, 47 Revue d'épidémiologie et de santé publique 7 (1999) ; S. Renault, Du concept de fragilité et de l'efficacité de la grille AGGIR, Gérontologie et société, 2/2004, pp. 83-107.

dépendance de la personne, apporteront des informations utiles pour adapter le type d'aide aux besoins de la personne. Pour chacune des variables retenues dans la grille, l'observateur attribue l'une des trois notes suivantes:

1. fait seul, totalement, habituellement et correctement ;
2. fait partiellement, ou non habituellement, ou non correctement ;
3. ne fait pas.

A l'issue de l'observation, un calcul simple est opéré et, selon le résultat obtenu, la personne est rattachée à l'un des six groupes qui vont du moins autonome au plus autonome. Le groupe 1 comprend les personnes âgées confinées au lit ou au fauteuil, dont les fonctions mentales sont gravement altérées et qui nécessitent la présence continue d'intervenants. Le groupe 6 comprend les personnes ayant totalement conservé leur autonomie dans les actes de la vie courante.

Il pourrait être tentant d'adopter le principe d'une telle démarche pour évaluer l'autonomie d'un robot. Il faudrait évidemment adapter les critères pour apprécier les trois conditions requises afin de considérer qu'un robot est autonome. Mais, l'opérationnalisation de la grille s'avère pour le moins compliquée. Prenons par exemple, la question de l'apprentissage qui est au cœur des débats en matière d'intelligence artificielle. Comment évaluer la capacité d'apprentissage d'un robot?

La récente victoire du robot «AlphaGo» de Google sur le champion du monde Lee Sebol a été considérée par certains comme une avancée décisive dans la voie de l'autonomisation de la machine.[6] C'est grâce à sa capacité d'apprentis- sage, en jouant contre elle-même, qu'«AlphaGo» aurait atteint son niveau d'excellence dans sa spécialité.[7] Un transhumaniste enthousiaste comme Laurent Alexandre verrait sans doute dans «AlphaGo» une étape clé sur la voie de la singularité technologique et la confirmation qu'elle se produira bien en 2025. Nul doute qu'une évaluation menée dans cet esprit donnerait un 3 bien senti.

Mais, quelques jours après la victoire «d'AlphaGo», l'intelligent artificielle con- naissait un échec tout aussi retentissant : l'interruption prématurée du «chatbot Tay» produit par les équipes non moins talentueuses de Microsoft. Comme tout robot de ce type, Tay était

6 L. Alexandre, Faut-il interdire Google-AlphaGo?, Le Monde – Sciences, 15 Mars 2016, http://www.lemonde.fr/sciences/article/2016/03/15/faut-il-interdire-google-alphago_4882798_1650684.html.

7 H. Yu, AlphaGo and the Declining Advantage of Big Companies, Harvard Business Review, 24 Mars 2016, https://hbr.org/2016/03/alphago-and-the-declining-advantage-of-big-companies.

destiné à entretenir des simulacres de conversation sur Tweeter à partir de phrases prédéfinies et de mécanismes d'apprentissage du même type que ceux «d'AlphaGo».[8] Pourtant, en moins de quinze heures, l'adolescente un peu superficielle qui souhaitait initialement la généralisation de la journée des animaux domestiques s'était transformée par l'effet de ses capacités d'apprentissage en une harpie extrémiste, appelant à brûler certaines personnalités féministes et convaincue du bienfondé des diverses théories du complot relatives au 11 septembre. Consternées par la tournure prise par l'expérience, les équipes de Microsoft ont préféré tout arrêter[9] et ranger au moins temporairement Tay dans la même armoire que Frankenstein. Un évaluateur même magnanime n'aurait pas pu donner une note supérieure à 1 à la malheureuse créature.

Comment expliquer qu'au même moment, des machines conçues par des équipes disposant des mêmes compétences puissent parvenir à des résultats aussi différents? Que peut-on en déduire pour évaluer l'état d'avancement de l'intelligence artificielle? Peut-on en tirer une vision prospective dans le domaine particulier des robots militaires?

L'explication de cette ambivalence des capacités d'apprentissage de la machine nous semble au fond assez simple.

Le robot «AlphaGo» déploie son intelligence artificielle dans un monde que l'on pourrait qualifier de «compliqué». Dans le jeu de go, le cadre à l'intérieur duquel se déploie la capacité d'apprentissage de la machine est parfaitement délimité, connu et intangible. Le terrain du jeu est connu et inscrit dans un espace invariable: dix-neuf cases par dix-neuf cases. Les règles du jeu sont codifiées, respectées par les joueurs et leur violation est immédiatement sanctionnée par un juge : deux joueurs et deux seulement, des pions de deux couleurs et deux seulement, une alternance régulière des coups joués par l'un et par l'autre, des mouvements autorisés et des mouvements interdits... Même si la stratégie d'apprentissage est plus riche que celle du Tic-Tac-Toe que la machine a maîtrisé dès les années 1960, même si elle est différente de la stratégie d'apprentissage pour le jeu d'échec où le champion du monde a été battu dans les années 1990,[10] «AlphaGo» intervient dans un environnement connu et intangible, strictement régulé où l'incertitude et les frictions n'existent pas, où l'adversaire n'ajoute pas de pions d'une troisième couleur ou ne renverse pas le plateau s'il est en difficulté. L'environnement est certes compliqué, mais il n'est que compliqué.

8 J. Huang et al., Extracting Chatbot Knowledge from Online Discussion Forums, IJCAI 2007.

9 M. Ingram, Microsoft's Chat Bot Was Fun for Awhile, Until It Turned into a Racist, Fortune, 24 Mars 2016, http://fortune.com/2016/03/24/chat-bot-racism/?iid=sr-link1.

10 M. Campbell et al., Deep Blue, 134 Artificial Intelligence 57 (2002); F.-H. Hsu, IBM's Deep Blue Chess Grandmaster Chips, IEEE Micro, 2/1999, pp. 70-81.

Le défi pour le robot «Tay» était beaucoup plus difficile puisque son intelligence devait se déployer dans un univers rendu complexe par le fait que les individus ave lesquels Tay entrait en relation n'étaient pas nécessairement animés d'intentions louables ou inoffensives[11] et qu'ils étaient prêts à «renverser le plateau» pour déstabiliser la machine et lui faire prendre des positions pour le moins inappropriées. Il n'a fallu que quinze heures à ces adversaires «irréguliers» pour par- venir à leurs fins et provoquer l'arrêt prématuré de l'expérience.

Qu'en déduire pour le cas des robots militaires si ce n'est que toute l'histoire des conflits nous montre que l'irrégularité est la règle et que les forces armées de tous les pays et de tous les temps ne se bornent pas à conduire leurs opérations en respectant un cadre prédéfini et connu de tous. Au contraire, les frictions et les surprises sont de tous les instants sur le champ de bataille.[12] Le monde de la guerre est par nature celui de l'incertitude et de la complexité. Il est donc celui de «Tay» et non celui «d'AlphaGo».

En conséquence, il est permis de penser que, malgré les progrès spectaculaires enregistrés par l'intelligence artificielle dans les vingt dernières années, la machine ne dispose pas aujourd'hui d'une capacité d'apprentissage suffisamment développée[13] pour lui donner une quelconque autonomie sur un champ de bataille. Si l'on considère que la différence entre un univers compliqué, tel celui du jeu de go, et un univers complexe, tel que celui du champ de bataille, n'est pas seulement une différence de degré mais une différence de nature, c'est à dire s'il n'y a pas de progression incrémentale de l 'un à l'autre mais une véritable rupture, il est permis de penser que la technologie capable de traiter de manière appropriée la complexité du champ de bataille n'est pas contenue en germe dans celle qui permet de vaincre l'homme sur un plateau d'échec ou de go.

11 W. Audureau, Derrière les dérapages racistes de l'intelligence artificielle de Microsoft, une opération organisée, Le Monde – Pixel, 25 Mars 2016, http://www.lemonde.fr/pixels/article/2016/03/25/derriere-les-derapages-racistes-de-l-intelligence-artificielle-de-microsoft-une-operation-organisee_4890237_4408996.html.

12 V. Desportes, Comprendre la guerre, Paris 2001.

13 R. Waters, Microsoft's Tay Shows AI Has Come a Long Way but Not Far Enough, Financial Times, 24 Mars 2016, https://www.ft.com/content/3c8c3728-f1e5-11e5-9f20-c3a047354386.

L'autonomie des systèmes d'armes: les confusions sémantiques

A cette première difficulté méthodologique qui rend périlleuse toute mesure visant à dresser une carte de l'autonomie supposée des machines, s'en ajoute une deuxième qui tient davantage à la question des concepts mobilisés au cours de l'analyse. En particulier, le terme «autonomie» est fréquemment utilisé à tort pour qualifier des dispositifs qui relèvent en réalité de catégories tout à fait différentes. Il convient à cet égard de bien différencier trois concepts fondamentalement distincts. La notion d'autonomie proprement dite doit être réservée à la seule hypothèse d'absence totale de contrôle ou de supervision de l'homme sur les choix opérés par la machine. Une machine autonome serait en mesure de prendre des décisions selon des procédures qui n'auraient pas été entièrement déterminée par son concepteur mais acquises à la suite d'un auto-apprentissage indépendant, ce qui introduirait un véritable aléa dans la nature des actions entreprises et des effets produits. Dans cette hypothèse, l'homme disparaîtrait complètement de la boucle de décision. Il n'aurait plus aucun pou- voir de validation ou de veto sur ce que peut entreprendre la machine. En ce sens, l'autonomie est complète ou elle n'est pas.

L'autonomie doit donc bien évidemment être distinguée de la notion de «télé-opération». Dans le cadre de systèmes télé-opérés, l'éloignement plus ou moins grand de l'opérateur par rapport au système sur lequel il agit n'amoindrit nullement le contrôle qu'il peut exercer sur lui. L'homme reste dans la boucle. Il décide des actions entrepris- es par la machine depuis son poste de pilotage. En tout état de cause, éloignement n'équivaut nullement à autonomie, au contraire. Le pilote de l'engin télé-opéré, qu'il en soit à quelques mètres comme dans le cas du robot démineur ou à plusieurs milliers de kilomètres comme dans le cas d'un drone, en conserve le contrôle complet. Il est dans la boucle. Pour ce qui est des décisions de ciblage et de tir en particulier, celles-ci font l'objet d'une chaîne de commandement parfaitement définie et les responsabilités de chacun des acteurs sont parfaitement identifiées.[14] Le développement des sciences et des techniques de télé-opération ne sont donc ni une préfiguration ni un stade préalable de l'autonomie des machines. On ne saurait, a fortiori, les confondre sans dénaturer complètement les termes du débat.

14 R. Doaré, Les robots militaires terrestres: quelles responsabilités en droit interne?, R. Doaré et H. Hude (éd.), Les robots au cœur du champ de bataille, Paris 2011, pp. 119-126; C. Fragata, Les problématiques juridiques liées à l'introduction des robots militaires terrestres, R. Doaré et H. Hude (éd.), Les robots au cœur du champ de bataille, Paris 2011, pp. 119-126.

Troisième et dernier concept qui doit être également distingué de l'autonomie, l'automatisation de certaines fonctions opérationnelles qui est une caractéristique commune, voire banale, des systèmes d'armes contemporains. On retrouve cette automatisation, par exemple, dès les années 1960 dans la défense sol-air où l'association d'un calculateur et de radars permettait de repérer un avion ennemi à plus de cent kilomètres et de fournir à un lanceur HAWK les éléments de tir requis pour le traitement de la cible. On la retrouve également dans des programmes désormais révolus où un missile avait pour fonction de provoquer la réaction d'un système anti-missile et, en analysant le signal émis par ce dernier, de l'identifier et, éventuellement, de le détruire s'il correspondait à un ou plusieurs systèmes définis comme des cibles. En permettant ainsi l'automatisation de certaines fonctions opérationnelles, le progrès technique donne à ceux qui doivent prendre des décisions sur le champ de bataille des outils de plus en plus puissants d'aide à la prise de décision dans des situations opérationnelles de plus en plus complexes où ils sont confrontés à des contraintes humaines et temporelles de plus en plus fortes. A une complexité sans cesse accrue des environnements et des missions doit répondre une complexité au moins égale des systèmes de décision et de commandement. Mais, la complexité induite par l'automatisation des fonctions opérationnelles des systèmes militaires a pour ambition et pour effet, non pas de substituer la machine à l'homme dans le processus de décision, mais, au contraire, de renforcer le contrôle que ce dernier exerce sur les décisions relatives à l'emploi de la force. Libéré de tâches que la machine peut accomplir à sa place (collecter des données, les mettre en forme…) permet au combattant de se recentrer sur les décisions importantes, ce qui, in fine, renforce le contrôle humain là où il est le plus précieux et le plus irremplaçable.

Si l'on devait résumer ce qui sépare fondamentalement l'autonomie des deux autres notions, la télé-opération et l'automatisation de certaines fonctions opérationnelles, sous le critère du contrôle exercé par l'homme sur les systèmes d'armes, il serait possible de dire que les deux dernières ont pour fonction et pour effet d'aider le combattant à prendre les décisions appropriées dans un contexte de plus en plus difficile tandis que l'autonomie aurait pour fonction et pour effet de substituer la machine à l'homme dans la prise de décision sur le champ de bataille. Comme on le voit, aucune carte de l'autonomie ne saurait être construite sans que ne soient respectées les distinctions fondamentales entre les différentes notions que l'on assimile parfois de manière parfois inconsciente mais toujours abusive.

Une fois cette classification établie, qu'en est-il de la place respective de l'autonomie, de la télé-opération et de l'automatisation des fonctions opérationnelles au sein des programmes de R&D militaires? Sont-elles au cœur des priorités retenues par les responsables de ces programmes? Bénéficient-elles d'une priorité manifeste en termes de crédits budgétaires?

En nous appuyant exclusivement sur des sources ouvertes, nous avons retenu, pour illustrer le propos, deux programmes de la DARPA, que l'on peut considérer comme l'une des institutions les plus avancées en matière de robotisation du champ de bataille.

Le premier programme est désigné par l'acronyme «SXCT» (Squad X Core Technologies) Il vise à mettre à la disposition des unités débarquées un ensemble d'informations très complètes afin de les aider à mieux appréhender l'environnement des missions qui leur sont confiées.[15] Pour ce faire, le programme SXCT doit permettre à la fois d'accroître la précision des coups portés à l'adversaire dans un rayon de mille mètres, de conduire les opérations cyber de nature à neutraliser les communications ennemies, d'identifier les menaces aussi bien que l'emplacement des adversaires ou des alliés. La DARPA évoque à propos de ce programme un processus de «détection autonome des menaces» («Autonomous Threat Détection») Or, on voit bien que, en dépit de l'utilisation du qualificatif «Autonomous», il ne s'agit nullement d'un dispositif «autonome» mais de l'intégration de fonctions automatisées qui permettent aux forces armées de mieux appréhender leur environnement et de prendre les décisions les plus adaptées en connaissance de cause.[16] Le programme SXCT va donc à l'encontre de la logique du SALA. Loin de permettre à la machine de s'immiscer dans la prise de décision, le système apporte une capacité supplémentaire de contrôle au combattant qui doit user de la force sur le champ de bataille. La décision reste la sienne ; elle est simplement mieux éclairée par l'automatisation de certaines fonctions opérationnelles qu'il devait auparavant accomplir lui-même ou qui n'étaient pas réalisées. Il s'agit donc bien d'un renforcement du contrôle humain sur les décisions prises par recentrage de la capacité de décision du combattant sur ce qui est essentiel dans l'action de combat.

Dans un autre de ses programmes portant le nom de «CODE»,[17] la DARPA fait également mention d'une «autonomie collaborative» entre l'homme et la machine («Collaborative Autonomy») Elle serait même l'une des caractéristiques majeures de ce programme. Mais, le terme est, ici aussi, trompeur puisque cette autonomie collaborative est ici encore une forme d'automatisation de certaines fonctions opérationnelles. Il s'agit, en effet, pour le pilote d'un avion de chasse de mettre en œuvre un ensemble de drones qui assurent des fonctions spécifiques (transmission, identification de cibles, frappe, guerre électronique…)

15 Squad X Core Technologies Takes First Steps toward Improving Capabilities for Dismounted Soldiers and Marines, 12 Octobre 2015, http://www.darpa.mil/news-events/2015-12-10.

16 The "program aims to develop novel technologies that could be integrated into user-friendly systems that would extend squad awareness and engagement capabilities without imposing physical and cognitive burdens" ibid.

17 Collaborative Operations in Denied Environment (CODE), http://www.darpa.mil/program/collaborative-operations-in-denied-environment.

sous le contrôle permanent («supervision») de ce pilote.[18] Nulle substitution de la machine à l'homme dans le processus de décision mais une démultiplication de ses capacités d'action à travers un système automatisé dont le pilote con- serve l'entier contrôle.

A travers ces deux exemples, il apparaît que l'ambition affichée dans les pro- grammes d'armement les plus récents et les plus innovants n'est pas d'écarter l'homme de la boucle de décision et de conférer aux robots la responsabilité de décider à leur place et en totale autonomie, c'est à dire sans aucun contrôle humain sur les choix opérés par la machine. Il s'agit, au contraire, de faire bénéficier les militaires des pro- grès techniques qui leur permettront de décider et d'agir à bon escient dans des environnements conflictuels particulièrement exigeants et sous des contraintes physiques et cognitives particulièrement lourdes. La machine joue alors pleinement son rôle d'aide à la décision en prenant en charge les activités dangereuses, répétitives ou fastidieuses et en permettant ainsi le recentrage du contrôle humain sur les éléments clés du processus de décision.

Les systèmes d'armes létaux autonomes: une aporie politique, stratégique et militaire

A travers les deux exemples mentionnés, les programmes SXCT et CODE, les responsables des investissements en R&D semblent montrer une inclination pour des dispositifs techniques qui correspondent à une logique d'aide à la décision humaine et non pas de substitution de la machine à l'homme dans les systèmes d'armes futurs. Peut-on expliquer cette inclination par le seul fait des avancées encore inachevées de l'intelligence artificielle, ce qui signifierait que la priorité à l'autonomie des systèmes d'armes n'est qu'une question de temps et qu'elle adviendra inéluctablement au fur et à mesure des progrès de l'intelligence artificielle? Plus largement, l'état du progrès scientifique et technique est-il le seul facteur dimensionnant de la réflexion sur le SALA? Y a-t-il dans ce domaine une forme de déterminisme technologique implacable? Quid en particulier de l'intérêt militaire que pourrait présenter un système d'armes létal autonome pour la réussite des missions confiées aujourd'hui aux forces armées? Cette question de l'utilité militaire du SALA apparaît d'autant plus cruciale que la notion même d'autonomie d'un système d'armes pose problème au regard de la conception politique qui préside de nos jours à l'intervention des forces armées dans une situation de conflit. La doctrine de la France (que l'on retrouve sous une forme ou sous une autre chez ses alliés) postule que l'intervention militaire n'a de sens que si elle crée les conditions politiques requises pour le rétablissement d'une paix durable. La victoire militaire permise par l'intervention des forces

18 "CODE intends to focus in particular on developing and demonstrating improvements in collaborative autonomy – the capability of groups of UAS to work together under a single person's supervisory control", ibid.

armées n'est pas une fin en soi ; elle est un préalable nécessaire. Elle ne garantit pas à elle seule le succès politique et stratégique. Elle doit être déterminée par «un objectif stratégique, déduit du but politique, et qui forme le point de convergence des actions militaires.»[19] Le succès politique et stratégique résultera de la capacité des forces armées à favoriser par leur action

> *«l'établissement des conditions indispensables au succès stratégique en instaurant la sécurité nécessaire après avoir établi le contrôle du milieu. L'objectif des opérations, décrites selon un continuum – intervention, stabilisation et normalisation – est la restaura- tion des conditions de vie normale pour la population d'une entité politique réinsérée dans la communauté internationale.»* [20]

Dans cette conception dite «globale» de l'action des forces armées, l'intervention militaire ne peut réussir sans un contrôle étroit et permanent de l'emploi de la force, sans une minimisation et, en tout cas, une maîtrise constante du degré de coercition utilisé pour atteindre l'objectif militaire. Or, cette maîtrise de tous les instants et de tous les niveaux, que l'on illustre souvent par la figure bien connue du «caporal stratégique», n'est possible que si les militaires sont en mesure d'anticiper la nature et de contrôler les effets des décisions prises dans la conception et la conduite des opérations. Un système d'armes qui agirait en totale autonomie, sans aucun contrôle humain, serait, par définition, incontrôlable et imprévisible dans ses décisions et dans ses actions et cette imprévisibilité serait d'autant plus inacceptable qu'elle aurait des conséquences létales. La notion «d'autonomie» apparaît donc contradictoire avec l'impératif contrôle que les forces armées doivent conserver dans l'emploi de la force et sans lequel la victoire militaire n'a pas de sens stratégique ou politique.

Développer en priorité des systèmes d'armes visant à l'autonomie entrerait en contradiction directe avec la conception politique des opérations militaires. A l'inverse, les innovations qui permettent une meilleure appréhension des données complexes du champ de bataille et qui contribuent ainsi à un contrôle plus resserré dans l'emploi de la force par l'amélioration des processus de décision du chef militaire ou du combattant s'inscrivent pleinement dans la philosophie de l'action globale des forces armées. C'est pourquoi les systèmes télé-opérés ou ceux qui permettent d'automatiser certaines fonctions opérationnelles et dont l'effet commun est d'améliorer les conditions de la prise de décision sur le champ de bataille sont utiles et attendus tandis que les systèmes autonomes qui prendraient des décisions à la place des militaires ne le sont pas.

19 Centre de Doctrine d'Emploi des Forces, Tactique Générale, FT-02, Paris, Juillet 2008, http://www.cdef.terre.defense.gouv.fr/publications/doctrine-des-forces-terrestres-francaises/les-documents-fondateurs/ft-02, p. 16.

20 Ibid.

Conclusion

Le sentiment de vertige qui peut parfois saisir l'observateur des grandes avancées de la science et des technologies ne doit pas le conduire à perdre tout esprit critique ni renoncer à toute rigueur dans l'évaluation de ces progrès. Cartographier les développements de l'autonomie des machines peut être un bon exercice à cet égard. La construction d'une carte prospective peut, en effet, être le moment privilégié de traitement d'un certain nombre de difficultés et de poser les bases saines d'une approche partagée. Nous avons à cet égard souligné trois points qui nous paraissent devoir être tranchés avant de se lancer dans la collecte des données qui permettront de constituer la carte.

Deux de ces points concernent la définition de l'objet d'étude et les méthodes permettant de l'appréhender.

De quoi parle-t-on au juste? Le respect de la définition donnée des systèmes d'armes létaux autonomes doit permettre de ne pas tout confondre et de diriger l'étude sur ce qui relève véritablement d'une hypothétique autonomie des systèmes d'armes et non pas de systèmes d'armes télé-opérés ou comportant des fonctions opérationnelles automatisées, tous systèmes d'armes qui sont déjà présents sur les champs de bataille.[21]

Comment mesurer sans biais exagéré le degré d'avancement des techniques relatives à l'autonomie, en particulier les progrès de l'intelligence artificielle? Si, dans ses principes de construction, une grille de lecture peut être construite en prenant pour modèle les outils relatifs à l'autonomie des personnes physiques, l'opérationnalisation d'une telle grille dans le cas des robots soulève des difficultés substantielles.

Enfin, le troisième et dernier point porte plus fondamentalement sur l'attitude à l'égard du progrès des sciences et des techniques. Celui-ci suit -il des lois inéluctables comme la trop fameuse loi de Moore? Est-il parfaitement extrapolable et détermine-t-il à lui seul l'évolution des systèmes d'armes utilisés sur le champ de bataille? Nous ne souscrivons pas sur ce point à l'approche des scientifiques qui, comme Stephen Hawking, pensent que l'avènement des systèmes d'armes létaux autonomes est une fatalité inéluctable. Nous pensons au contraire que les choix technologiques s'inscrivent dans une perspective plus large, politique, stratégique, militaire économique… qui, dans les conditions actuelles, rend plus qu'improbable, et pour tout dire aporique la perspective de développer de tels systèmes.

21 D. Danet et al., La robotisation du champ de bataille: évolution ou robolution?, D. Danet et al. (éd.), La guerre robotisée, Paris 2012, pp. 5-28.

The Security Impact of Lethal Autonomous Weapons Systems

Jayantha Dhanapala*

Introduction

Throughout the passage of history, the pursuit of war has undergone profound transformations with the introduction of new weapons and strategies influenced by the application of new technology. The invention of gunpowder in China one thousand years ago has had a tremendous impact on the way wars are fought; we see the consequences daily in the use of firearms and explosive weapons. Nuclear weapons were another so-called technological advance in warfare. We are also still trying to deal with their possibly catastrophic humanitarian, ecological, and genetic consequences of its use since 1945, with political intent or by accident, and by state or non-state actors. Some have described lethal autonomous weapons systems as the third category of weapons to pose a grave danger to human security and to global peace and stability.[1]

As the founder of the World Economic Forum, Klaus Schwab, notes, the history of warfare and international security is the history of technological innovation, and today is no exception.[2] Modern conflicts involving states are increasingly 'hybrid' in nature, combining traditional battlefield techniques with elements previously associated with non-state actors. The distinction between war and peace, combatant and non-combatant, is becoming uncomfortably blurry, and this is profoundly impacting the nature of national and international security, affecting both the probability and the nature of conflict.

Over the past century, an elaborate tapestry of international law, regulations, and machinery has been woven that sets limits on the means and methods of warfare and specific weapons. To constrain armed conflict and law enforcement operations, it has established that the use of force should be an option of last resort. Similarly, practices have evolved aimed at pursuing the peaceful settlement of disputes through collective and co-operative security frameworks, some of them regional.

* Governing Board Member, Stockholm International Peace Research Institute.

1 The comparison in major revolutions in military affairs was first made by P. Singer, Wired for War, New York 2009, pp. 179 et seq., notably p. 203.

2 K. Schwab, The Fourth Industrial Revolution: What It Means, How to Respond, 14 January 2016, http://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond.

Due to these and other measures, one nation's security cannot be achieved by causing the insecurity of others. This requires that military expenditure and defence arsenals should be commensurate with the genuine needs of countries.

The CCW uniquely combines the humanitarian strand of international relations with disarmament and arms control. Since 2014, it has grappled with a broad range of concerns raised by the prospect of weapons that, once activated, would select and engage targets without human control. In this chapter, I will focus on the peace and security implications of lethal autonomous weapons systems.

Security concerns associated with lethal autonomous weapons systems

In the CCW deliberations to date, a number of countries including my own have elaborated particular concern at the possibility that lethal autonomous weapons systems will negatively impact peace and destabilise regional and global security.[3] The most commonly cited security concerns are that these weapons have the potential to:

1. escalate the pace of warfare and the likelihood of resorting to war, in large part due to the promise of significantly reduced military casualties;
2. ignite and foster arms races, in case the possession by some states but not others leads to all states feeling compelled to acquire them, against the possibility of asymmetric warfare as a result of the discrepancies between technological haves and have-nots;
3. be acquired and used by non-state armed groups, including terrorist entities;
4. undermine existing law, controls, and regulations.

Each of these key concerns will be briefly examined, starting with the ever-increasing pace of warfare.

3 Statement of Sri Lanka, CCW Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, 13 April 2015, http://www.unog.ch/80256EDD006B8954/(httpAssets)/30534E70A6CFAAC-6C1257E26005F2B19/$file/2015_LAWS_MX_Sri+Lanka.pdf.

Escalate the pace of war and the likelihood of going to war

As the chair's report of the 2015 CCW experts meeting notes, one reason for military interest in autonomous functions is the increasing speed or pace of warfare.[4] This is despite warnings from scientists and others about the risks posed by autonomous systems interacting at speeds beyond human capacities.

A 2013 statement endorsed by more than 270 engineers, computing and artificial intelligence experts, roboticists, and professionals from related disciplines in 37 countries, asked how devices made and deployed by opposing forces and controlled by complex algorithms will interact, warning they could "create unstable and unpredictable behaviour (...) that could initiate or escalate conflicts, or cause unjustifiable harm to civilian populations".[5]

We are familiar with human error in the handling of technology with catastrophic consequences from tragedies such as Three Mile Island, Chernobyl, Fukushima, and Bhopal. Nuclear warheads accidentally have dropped off planes by developed countries with the most sophisticated safety systems in place. Thus, accidents from fully autonomous weapons cannot be ruled out.

A new report by the Center for a New American Security on the operational risks associated with autonomous weapons looks at potential types of failures that might occur in completely automated systems, as opposed to the way such weapons are intended to work.[6] The report finds they could be uncontrollable in real-world environments where they would be subject to design failure as well as hacking, spoofing, and manipulation by adversaries. The consequences of a failure that causes a weapon to engage an inappropriate target could be far greater with an autonomous weapon, resulting in "fratricide, civilian casualties, or unintended escalation in a crisis". It finds that autonomous weapons have a qualitatively different degree of risk than equivalent semi-autonomous weapons that would retain a human in the loop.

4 Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Submitted by the Chairperson, CCW/MSP/2015/3, 2 June 2015. http://www.unog.ch/80256EDD006B8954/(httpAssets)/587A415BEF5CA08BC1257EE0005808FE/$file/CCW+MSP+2015-03+E.pdf.

5 Campaign to Stop Killer Robots, Scientists Call for a Ban, 16 October 2013, http://www.stopkillerrobots.org/2013/10/scientists-call/.

6 The report's author is P. Scharre, who served as a U.S. Army Ranger in Iraq and Afghanistan, and who co-authored the 2012 Pentagon Policy Directive on autonomous weapons; see P. Scharre, Autonomous Weapons and Operational Risk, Center for a New American Security, February 2016, https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf, p. 12.

Weapons that have no means of human intervention after initial manufacture and programming – in terms of the use of judgment, discretion, or reasoning – are inherently insecure. The hacking of confidential communications through Wikileaks and other means has so far resulted in acute embarrassment in inter-state relations, but that could be far worse with the introduction of lethal autonomous weapons systems.

In 2014, more than 20 Nobel Peace Laureates including Pugwash issued a joint statement calling for a pre-emptive ban on fully autonomous weapons, which expressed concern that "leaving the killing to machines might make going to war easier and shift the burden of armed conflict onto civilians".[7]

Human Rights Watch and others have described "insurmountable legal and practical obstacles" that would likely interfere with holding someone accountable for unforeseeable, unlawful acts committed by a fully autonomous weapon.

At the 2015 experts meeting, the Holy See shared a 10-page paper exploring fundamental ethical questions relating to the use of fully autonomous weapons. It found that a lack of accountability could promote the use of fully autonomous weapons "because of the impunity they permit". It concluded, "the risks of deresponsibilization, dehumanization, and depoliticization induced by the use of lethal weapons removed from effective control by men are important enough that we can envisage asking for their prohibition".[8]

Proliferation and asymmetric warfare

The longer lethal autonomous weapons systems go unregulated, the greater the risk of their proliferation, especially to non-state actors. This problem is aggravated further when and if such weapons are considered having essential military benefits.

7 Signatories include former presidents Lech Walesa of Poland, Oscar Arias Sánchez of Costa Rica, F.W. de Klerk of South Africa, and José Ramos-Horta of Timor-Leste. Other individual signatories include Jody Williams, Mairead Maguire, Betty Williams, Rigoberta Menchú Tum, Shirin Ebadi, Leymah Gbowee, and Tawakkol Karman, who are members of the Nobel Women's Initiative, a co-founder of the Campaign to Stop Killer Robots. Nobel Women's Initiative, Nobel Peace Laureates Call for Preemptive Ban on 'Killer Robots', 12 May 2014, http://nobelwomensinitiative.org/nobel-peace-laureates-call-for-preemptive-ban-on-killer-robots/?ref=204.

8 The Holy See, The Use of Lethal Autonomous Weapons Systems: Ethical Questions, CCW Experts Meeting, 16 April 2015, http://www.unog.ch/80256EDD006B8954/(httpAssets)/4D28AF2B8BBBECED-C1257E290046B73F/$file/2015_LAWS_MX_Holy+See.pdf.

Who are the 'haves' with autonomous weapons technology? An increasing number of states are actively pursuing precursors to fully autonomous weapons and not all are transparent about their plans. Initial research and development has been identified as taking place in at least six countries.[9]

The 2015 open letter calling for a ban, that more than 3,000 artificial intelligence and robotics experts signed, affirms that "the key question for humanity today is whether to start a global AI arms race or to prevent it from starting". Its signatories find that autonomous weapons may not require "costly or hard-to-obtain raw materials", making them likely to become "ubiquitous and cheap for all significant military powers to mass-produce" and to "appear on the black market" and in the hands of terrorists, dictators, and warlords.

To quote further: "[i]f any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kaleshnikovs of tomorrow (...). Autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group." [10]

The risk of proliferation is far greater now than at any other time in the past, as technology is widely shared for commercial as well as military purposes. An arms race that confers advantages on one side over the other is harmful to the common security of humankind. The temptation to use technology already developed and incorporated into military arsenals would be great, and countries would be reluctant to give it up, especially if their competitors were deploying it.

As one delegation noted at the last CCW experts meeting in 2015, the use of lethal autonomous weapons could change not only the way war is fought, but how wars end.[11] If one resorts to lethal autonomous weapons systems, what about terminating the conflict? When do you stop? When is a political objective achieved?

Battlefield decisions taken by machines on auto-pilot can thrust nations into wars that they did not anticipate or desire. We can envisage many scenarios whereby a lethal autonomous weapon could sabotage peace deals and gravely endanger ceasefires concluded after pains-

9 The U.S., China, Israel, South Korea, Russia, and the UK.

10 Campaign to Stop Killer Robots, Artificial Intelligence Experts Call for Ban, 28 July 2015 http://www.stopkillerrobots.org/2015/07/aicall/.

11 Human Rights Watch, Losing Humanity: The Case against Killer Robots, November 2012, http://www.hrw.org/reports/2012/11/19/losing-humanity-0.

taking negotiations, for example when they are programmed to activate without detection after such agreements have been concluded. A normal weapon has a measurable time span between its launch and its impact, while a lethal autonomous weapon may not be constrained by any time span after its launch, making its impact timeless.

Non-state armed groups

The Middle East, Asia, and other regions are experiencing growing political extremism from nationalist groups, ethno-religious movements, and other non-state actors, for whom international norms are irrelevant. The technology is increasingly within reach with, as one research group puts it, "ever-more advanced drones capable of carrying sophisticated imaging equipment and significant payloads (...) readily available to the civilian market".[12]

The chair's report of the 2015 experts meeting notes how in future armed conflicts, "tactical considerations will require systems to be small, durable, distributable and stealthy", and "these developments could lead to an increased risk of an arms race and proliferation, as smaller systems are more easily acquired by nonstate actors".[13] It describes how lethal autonomous weapons systems "would be attractive to non-state actors, enabling them to create shock and awe, to use them as a force multiplier, and to spare their own fighters".[14]

Undermine existing law

The introduction and application of lethal autonomous weapons systems to the battlefield has frightening implications for the laws of war, especially proportionality, precaution in attack in the context of other options, accountability, and the important distinction between combatant and civilian, which a programmed robot cannot yet discern. The weapons have the potential to weaken the role and rule of international law and undermine the international security system in the process.

12 Remote Control Project, Hostile Drones: The Hostile Use of Drones by Non-State Armed Actors Against British Targets, January 2016, http://remotecontrolproject.org/wp-content/uploads/2016/01/Hostile-use-of-drones-report_open-briefing.pdf.

13 Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (n 4).

14 Ibid.

As Special Rapporteur Christof Heyns reminded us in his 2013 report, these weapons "could have far-reaching effects on societal values, including fundamentally on the protection and the value of life", as they would be unlikely to possess qualities necessary to comply with existing international humanitarian law, such as "human judgment, common sense, appreciation of the larger picture, understanding of the intentions behind people's actions, and understanding of values and anticipation of the direction in which events are unfolding".[15] The inability of fully autonomous weapons to interpret intentions and emotions would be a significant obstacle to compliance with the rule of distinction.

Heyns identified "built-in constraints that humans have against going to war or otherwise using force" that "continue to play an important (if often not decisive) role in safeguarding lives and international security", particularly "unique human traits such as our aversion to getting killed, losing loved ones, or having to kill other people".[16]

As it is, human discernment and judgment have been known to be severely impaired or limited by the 'fog of war'. A machine will be thrown into a dysfunctional state.

International human rights law considerations also apply at all times to lethal autonomous weapons systems and cannot afford to be ignored, including their potential impacts on the right to life, the right to bodily integrity, the right to human dignity, the right to humane treatment, and the right to remedy. This year's report to the Human Rights Council on the proper management of peaceful assemblies finds that "autonomous weapons systems that require no meaningful human control should be prohibited".[17]

The precautionary principle is also relevant to these deliberations.[18] So is the fundamental Martens Clause, which mandates that the 'principles of humanity' and 'dictates of public conscience' be factored into an analysis of their legality.

15 C. Heyns, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions for the Human Rights Council, A/HRC/23/47, 9 April 2013, http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf.

16 Ibid., para. 57.

17 Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns, and Special Rapporteur on the Rights to Freedom of Peaceful Assembly and of Association, Maina Kiai, for the Office of the High Commissioner for Human Rights, A/HRC/31/66, 4 February 2016, https://t.co/hpkjz7CfyV.

18 Principle 15 of the Rio Declaration on Environment and Development, adopted 14 June 1992, U.N. Doc. A/CONF.151/26 (vol. 1), 31 ILM 874, 1992, http://www.un.org/geninfo/bp/enviro.html.

Ways to deal with security concerns

The forum provided by the CCW has made a significant start in deliberating the challenges posed by lethal autonomous weapons systems, but its slow pace and lack of a goal has seen it criticised for treading water.[19] Some view the CCW process as prolonging the debate without results while the autonomous systems are weaponised.

A basic flaw in our current approach seems to be the failure to consider the challenges of lethal autonomous weapons systems from the perspective or point of view of a potential human victim, especially non-combatant civilians. Sierra Leone is one of few states asking about the implications of a machine being able to take human life.[20] As the World Council of Churches observed at the first CCW meeting on this topic in 2014, "we have heard very little from states who would not be able to acquire these weapons, who are on the receiving end, and already know how these weapons will look and feel".[21]

You do not have to imagine a specific country where autonomous weapons could be used. Just look at what is happening in the world today, and the scenarios unfortunately come easily. What happens when autonomous attack aircraft replace human drone pilots altogether? Or when police deploy swarms of small drones equipped with less-than-lethal weapons to monitor mass protests? Or when stationary autonomous weapons with sensors are placed along long borders?

Who will be the victims of those weapons? Will they be groups of men and women who have been deemed enemy combatants? Democracy fighters and social justice campaigners? Economic migrants and asylum seekers?

19 Statement by the Campaign to Stop Killer Robots, Convention on Conventional Weapons Annual Meeting of High Contracting Parties, Geneva, 13 November 2015, http://www.unog.ch/80256EDD006B8954/(httpAssets)/192F54E302628EDFC1257F0F005EBAD4/$file/2015_CCWMSP_LAWS_Human+Rights+Watch.pdf.

20 Statement by Sierra Leone, CCW Informal Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, 12-16 May 2014.

21 See Campaign to Stop Killer Robots, Report on Activities, Convention on Conventional Weapons second informal meeting of experts on lethal autonomous weapons systems, Geneva, 13-17 April 2015, http://www.stopkillerrobots.org/wp-content/uploads/2013/03/KRC_CCWx2015_Report_4June2015_uploaded.pdf, p. 19.

The conflicts burning today in Syria, Yemen, and other countries show how the 21st century battlefield is never completely clear or 'clean' of civilians. There's no way to guarantee or ensure that the civilian toll will diminish with the introduction of lethal autonomous weapons systems. Saving soldiers' lives by not placing them at risk means placing civilian lives in harm's way.

The concept of meaningful, appropriate, or adequate human control over the critical targeting and attack functions in weapons systems and their use in individual attacks has swiftly acquired currency within this international debate. Many states have affirmed the need to retain human control of autonomous weapons and want to further explore this concept as an approach to tackling future weapons systems.[22]

Last October, Germany affirmed that the deliberations on autonomous weapons to date show that "there is a common understanding that machines should not be allowed to take life-and-death decisions without human intervention".[23] The "rejection of fully autonomous weapons systems deciding over the use of force against humans without any human intervention" is a key common understanding listed by the 2015 meeting chair's report.[24]

Conclusion

Decisions on war and peace are political and security issues requiring serious debate and discussion. By considering these weapons from the perspective of human control we can begin to assert some authority over our future. It is our collective moral and social responsibility to ensure that international law protects humankind.

Our community has a problem with timeliness and responsiveness, as evidenced by the moribund Conference on Disarmament. The CCW proved its relevance by swiftly adopting a discussion mandate on this topic, but there has been little progress beyond carefully and

22 Austria, Croatia, Germany, Iraq, Ireland, Japan, Mexico, the Netherlands, Poland, Russia, South Africa, and Zimbabwe requested that meaningful human control be part of the deliberations in 2016. Belgium, Colombia, Sweden, and the ICRC affirmed the importance of human control of weapons systems.

23 Statement of Germany, UNGA First Committee on Disarmament and International Security, New York, 9 October 2015, http://reachingcriticalwill.org/images/documents/Disarmament-fora/1com/1com15/statements/9October_Germany.pdf.

24 Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (n 4).

objectively building a base of common knowledge. That is because we lack a goal to work towards. If this challenge is to remain in the hands of CCW states, states must feel compelled to take bold steps and avoid a disappointing, inadequate response or none at all.

Expectations are riding high. Procrastination is no recipe for action, and low expectations make poor standards. No one wants a long, drawn-out, and inconclusive process. Policymakers should focus on bold measures that can take the CCW swiftly forward towards a solid and lasting outcome.

In closing, we should remind ourselves of the conclusion of the London Manifesto confronting a future of weapons of mass destruction, co-authored by Albert Einstein and the first Pugwash president Lord Bertrand Russell in 1955:

> *"There lies before us, if we choose, continual progress in happiness, knowledge, and wisdom. Shall we, instead, choose death, because we cannot forget our quarrels? We appeal, as human beings, to human beings: Remember your humanity, and forget the rest."*[25]

25 The Russell-Einstein Manifesto, issued in London, 9 July 1955, http://www.umich.edu/~pugwash/Manifesto.html.

Human Control in the Targeting Process

Merel Ekelhof*

Introduction

In current practice, there is an implicit expectation that humans exercise some form of control over the use of weapon systems. The term 'autonomy' in autonomous weapon systems, however, suggests that humans are not in control of these weapons. This common narrative is somewhat misleading as it fails to take into account the practical realities of targeting, in particular the adherence to a targeting process. This article focuses on the context of the *use* of autonomous weapons by offering an analysis of this targeting process, insofar as it is relevant to the debate over autonomous weapons. As these systems do not operate in a vacuum, it is necessary to offer some context regarding targeting and the use of weapon systems by the military.

The types of weapon systems under discussion here are Lethal Autonomous Weapons Systems (LAWS). The language to accurately describe these weapons is still unsettled. Whereas some describe LAWS as future weapons and not as existing weapons,[1] others argue that current and emerging weapon systems are both to be included in the debate on LAWS.[2] Nevertheless, regardless of what definition one decides to use, the contextual approach that focuses on the decision-making processes in which humans exercise control remains important in any discourse considering autonomous weapons. In addition, irrespective of what position one chooses – whether LAWS should be banned or not – the debate on these systems should include discussions about these processes. As such, I will

* LL.M; PhD researcher at the Faculty of Law (Transnational Legal Studies), Boundaries of Law, VU University of Amsterdam; and a Research Fellow at the Centre of the Politics of Transnational Law. Her research focuses on the development and use of autonomous weapon systems, with a particular focus on human control within relevant decision-making processes. The research is funded by the Netherlands Ministries of Defense and Foreign Affairs. The views expressed herein are those of the author and should not be construed as the position of either the Netherlands Ministry of Foreign Affairs or the Ministry of Defense. Email: m.a.c.ekelhof@vu.nl.

1 See e.g. the U.S. and the United Kingdom during the UN CCW Meeting of Experts General Exchange on 11 April 2016 in Geneva, Switzerland. See also the Non Paper on Characterization of a LAWS by France, presented during the CCW Meeting of Experts on LAWS in Geneva, 11-15 April 2016.

2 See e.g. the ICRC during the UN CCW Meeting of Experts General Exchange on 12 April 2016 in Geneva, Switzerland. See also the Working paper presented by Switzerland, 'Towards a 'compliance-based' approach to LAWS', during the Meeting of Experts on LAWS in Geneva, 11-15 April 2016.

not propose a working definition to describe Lethal Autonomous Weapons Systems. I will offer, however, an approach to the concept of 'Meaningful Human Control' by considering the context in which LAWS operate.

The targeting process

The 'loop' has become a very familiar term within the debate about the use of autonomous weapons. Generally, the loop is explained as comprising three distinct categories: weapons with a human 'in the loop', weapons with a human 'on the loop', and weapons with a human 'out of the loop'. To have a human 'in the loop' is commonly described as the capability of a machine to take some action, but subsequently stop and wait for a human to take a positive action before continuing. Being 'on the loop' means that humans have supervisory control, remaining able to intervene and stop a machine's on-going operation. The third modus of machine autonomy, 'out of the loop', is usually defined as the machine's capability of executing a task or mission without human intervention or even the possibility to intervene.[3]

The advantage of using this model to describe autonomy in weapon systems is its focus on the human-machine interface. It appears to be a useful method as it is easier for observers to relate to the described role as human operators or supervisors than to be forced to conceive the issue in merely abstract terms, which is crucial in relation to a topic as controversial as machine autonomy.[4] Nevertheless, it is not always clear what exactly is meant when discussing the 'loop'. According to Singer, there currently is a shift occurring that aims to redefine the meaning of having a human 'in the loop'.[5] In connection to this, Kurzweil

3 P. Scharre, Autonomous Weapons and Operational Risk, Center for a New American Security, Washington 2016, p. 9; Human Rights Watch and International Human Rights Clinic, Losing Humanity – The Case against Killer Robots, 2012, p. 2; U.S. Department of Defense, Defense Science Board, Task Force Report: The Role of Autonomy in DoD Systems, 2012; M.N. Schmitt and J.S. Thurnher, 'Out of the Loop': Autonomous Weapon Systems and LOAC, 4 Harvard National Security Journal 231 (2013); International Committee of the Red Cross, Report of the Expert Meeting on Autonomous Weapon Systems, 2014, p. 14.

4 The word 'autonomy' can mean a number of things. David Woods et al. explain that "[t]here is a myth that autonomy is some single thing and that everyone understands what it is". However, the word autonomy is employed with different meanings and intentions and can be viewed from many different angles. For example, autonomy, as described by the Stanford Encyclopedia of Philosophy, refers to self-governance, the capacity to be one's own person and to live one's life according to reasons and motives that are taken as one's own. However, technologically, autonomy means no more than the capability for unsupervised operation; see J.M. Bradshaw, R.R. Hoffman, M. Johnson, and D.D. Woods, The Seven Deadly Myths of 'Autonomous Systems', IEEE Intelligent Systems 2013, p. 2.

5 P.W. Singer, Wired for War, New York 2009, p. 125.

argues that 'in the loop' is becoming no more than "a political description".[6] Furthermore, Marra and McNeil claim that the debate over whether humans are 'in the loop' or 'out of the loop' has an all-or-nothing feel, and does not adequately account for the complexity of some technologies.[7] Clearly, it is not always straightforward what is meant by having a human in, on, or out of the loop. I propose to conceive the loop as the targeting process which is used by the military to plan, execute and assess military missions.[8] Specifically, NATO's targeting process can serve as an example of how weapons are used and how humans can exercise control over increasingly autonomous weapon systems.[9]

The term 'targeting' is often associated with the actual use of force, i.e. the lethal attack, the action that triggers kinetic energy (e.g. the firing of a weapon at a target). In other words, many associate targeting with something that looks like a destructive and lethal combination of power and strength.[10] However, the targeting process entails more than the actual unleashing of kinetic energy; there is, as the name implies, an entire process or decision-making cycle that precedes this moment. The targeting process is an iterative process that aims to achieve mission objectives in accordance with the applicable law and rules of engagement (ROE) by means of the thorough and careful execution of six phases. It is a process that has emerged over the course of history and has become embedded in the training and execution of NATO's military operations.[11] Furthermore, the targeting process can be found in both NATO doctrine and national military doctrines.[12] NATO doctrine describes joint targeting as "the process of determining the effects necessary to achieve the commander's objectives, identifying the actions necessary to create the desired effects based on means available, selecting and prioritizing targets, and the synchronization of fires with other military capabilities and then assessing their cumulative effectiveness and taking remedial action if necessary."[13]

6 Ibid.
7 M.C. Marra and S.K. McNeil, Understanding 'The Loop': Regulating the Next Generation of War Machines, 36 Harvard Journal of Law & Public Policy 1185 (2014).
8 See also the advisory report of the Joint Committee on AWS, Autonomous Weapon Systems – The need for Meaningful Human Control, No. 97 AIV / No. 26 CAVV, 2015.
9 NATO Allied Joint Publication, AJP-3.9 Allied Joint Doctrine for Joint Targeting, 2008.
10 G. Di Marzio, The Targeting Process...This Unknown Process (Part I), NRDC-ITA Magazine, no. 13, 2009, p. 11.
11 M. Roorda, NATO's Targeting Process: Ensuring Human Control Over and Lawful Use of 'Autonomous' Weapons, NATO Allied Command Transformation Publication on Autonomous Systems, 2015, p. 155.
12 See e.g. Ministerie van Defensie, Joint Doctrine Publicatie 5 – Commandovoering, 2012; NATO Allied Joint Publication, 2008.
13 The term 'joint' refers to the joint effort between all armed force components; see NATO Allied Joint Publication, 2008.

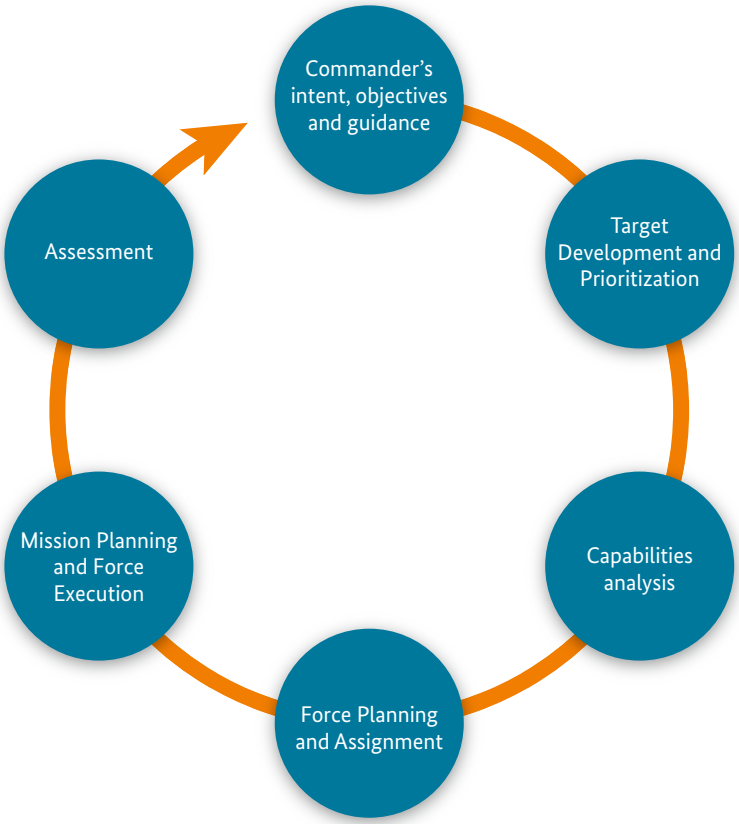
Within the process, there are two approaches to targeting: deliberate and dynamic. Deliberate targeting concerns the pursuit of previously selected targets which are known to exist in the operational area with lethal or non-lethal actions scheduled against them.[14] In other words, deliberate targeting is planned ahead of time. Dynamic targeting, by contrast, pursuits targets known to exist in the area of operation, but which were not detected, located or selected for action in time in order to be included in the deliberate process.[15] These targets may, for example, be fleeting. The steps taken in dynamic targeting are largely the same as those during deliberate targeting; the two approaches mostly differ in the speed with which steps are executed.[16]

14 NATO Allied Joint Publication, 2008, at 1.2.

15 Ibid.

16 B. Kastan, Autonomous Weapons Systems: A Coming Legal 'Singularity'?, 2013:1 Journal of Law, Technology & Policy 58 (2013). The time with which steps are executed in the targeting process is closely connected to the debate about increasingly autonomous weapon systems. As the dynamic process emphasises speed and is more reactive, it is not always clear how meaningful human control would be exercised in the dynamic targeting process if autonomous weapon systems executed critical functions of target selection and attack independent of human control. This raises questions that deserve attention. However, for the purposes of this short article, I will limit the analysis to the deliberate targeting process and will not focus on additional questions that are specifically raised by dynamic targeting or time-sensitive targeting.

The diagram is an illustrative example of the different steps in a targeting process. It is applicable to both the deliberate and dynamic approaches. It is an oversimplification – the targeting process is not a clear linear process, it requires constant feedback and reintegration in different phases – but it offers a useful lens for understanding the context within which weapon systems that have autonomy in their critical functions operate.



NATO doctrine[17] describes the six phases of the 'Joint Targeting Cycle' as follows:

Phase 1: commander's intent, objectives and guidance

The commander must clearly identify what to accomplish, under what circumstances and within which parameters. During this first phase, the goal of the mission is identified and analysed. Attainment of clear and achievable objectives is essential for the successful realisation of the desired end state.[18]

Phase 2: target development, validation, nomination and prioritisation

This phase aims to answer the question of what targets need to be effected to achieve the goal? Target development aims to identify different eligible targets that can be influenced. In this phase, the target validation ensures compliance with relevant international law and the rules of engagement.[19] For example, the principle of distinction plays a vital role in this phase to ensure that offensive action is only directed against military objectives and combatants, making a clear distinction between them and civilian objects and civilians.[20] In addition, during target development issues related to collateral damage may also become apparent and must be considered.[21]

Phase 3: capabilities analysis

Capabilities are analysed to assess what methods and means are available and most appropriate to generate the desired effects. Collateral Damage Estimation (CDE), which started in phase 2, remains a critical component of the analysis.[22]

17 NATO Allied Joint Publication, 2008.
18 U.S. Joint Chiefs of Staff, Joint Publication 3-60, Joint Targeting, II-4, 2013.
19 Specific IHL rules cannot be categorised according to these phases and often play a role in several of them. At the very least, the end result of the process must comply with all applicable law, see Joint Committee on AWS, Autonomous Weapon Systems – The need for Meaningful Human Control, No. 97 AIV / No. 26 CAVV, 2015.
20 Art. 48, 51 and 52 of Additional Protocol I.
21 No attack is to be launched that is expected to cause collateral damage excessive to the concrete and direct military advantage anticipated; see Article 51(5)(b) and 57(2)(iii) of Additional Protocol I.
22 Roorda (n 12), p. 159.

Phase 4: commander's decision, force planning and assignment

During force planning and assignment, capabilities are matched to the targets. This phase integrates output from phase 3 with any further operational considerations. After final approval is given, this information will be passed on to the assigned unit.

Phase 5: mission planning and force execution

This stage deals directly with planning and execution of tactical activity. The assigned unit will take similar steps as in phases 1 to 4, but on a more detailed, tactical level. Assessments in this phase take into account operational and legal standards, including the obligation to take feasible precautions in attack.[23] And, importantly, there is force execution during which the weapon is activated, launched, fired, or used (or refrained from use).

Phase 6: assessment

The combat assessment is performed to determine whether desired effects have been achieved. This feeds back into phase 1, and goals and tasks can be adjusted accordingly.[24]

Human control in the targeting process

Although autonomous weapons are often described as weapons that do not yet exist, there are some current examples of existing weapons with autonomy in the critical functions of target selection and attack. These can be certain types of missile and rocket defence weapons (e.g. the US Phalanx Close-in Weapon System), vehicle 'active protection' weapons (e.g. the Israeli Trophy), sentry weapons (e.g. the South Korean Aegis I & II) and loitering munitions (e.g. the Israeli Harpy). It can be relevant to include these systems in the analysis because it helps us gain a better understanding of how autonomy is already used and where problems could arise when the boundaries towards more autonomy are being pushed further.

23 Precautionary measures include: doing everything feasible to ensure the target is a lawful military target, taking all feasible precautions in the choice of means and methods of attack with a view to avoid or minimise collateral damage, cancelling or suspending an attack if it becomes apparent that the target is not a lawful military objective or if the attack will be disproportionate, giving effective warning if the circumstances permit (Article 57 Additional Protocol I); see ibid., p. 160.

24 NATO Allied Joint Publication, 2008., at 2.1-2.4.

Current examples of autonomous weapons are weapons that are activated by humans in phase 5 of the targeting process (force execution). After activation, there is an inevitable moment after which humans can no longer influence the direct effects of the use of force.[25] This is for example the case with the Israeli Harpy, which is programmed to select[26] and engage hostile radar signals in a predefined area. The Harpy seeks out and identifies an adversary's radar emitter by comparing the signal to its library of hostile emitters. Once the target is verified, the Harpy will attack it by detonating its warhead just above the target.[27] After activation, humans can no longer intervene in the process of target selection and attack. However, does that mean that humans are not in control of the autonomous weapon? Looking at the targeting process, it becomes clear that although parts of the mission will be executed by the weapon system autonomously, the targeting process as a whole is still largely human-dominated. Before an autonomous weapon system is deployed to conduct its assigned tasks in phase 5, humans have carried out an extensive planning stage in which humans formulate the overall goals, gather intelligence, select and develop targets, analyse the most suitable weapon, and decide under what circumstances and preconditions to employ a particular weapon that is in accordance with the relevant international law and ROE. Thus, even though an autonomous weapon selects and attacks a target in phase 5, it is not truly autonomous in the overall targeting process. It seems that, through this process, humans can remain in control of an autonomous weapon's actions on the battlefield, even though there is no direct human control over the system's critical functions of target selection and attack.

Key elements or components of Meaningful Human Control

'Meaningful Human Control' is often described by pointing out key elements or components of human control. The UK-based not-for-profit organisation Article 36 explains that humans can exercise control by assigning operational constraints by programming a predefined geographical area in which the autonomous weapon system has to limit its operation, and by limiting the time within which the system is allowed to operate without

25 Roorda (n 12), p. 165.

26 Current examples of weapons with autonomy in the critical functions of target selection and attack are generally systems that are pre-programmed to select a particular category of targets (e.g. hostile radar emitters). This type of target selection does not include a deliberate process in which the system itself analyses the lawfulness of the engagement. Hence, this type of target selection is not conducted in phase 2 of the targeting process (target development and validation), but rather in phase 5 (force execution). In current examples of weapons with autonomy in their critical functions, target selection can generally be understood as target recognition, rather than target development or mission planning.

27 See 'Harpy Air Defense Suppression System', Defense Update, https://defense-update.com/directory/harpy.htm.

direct human control.[28] The Center for a New American Security explains that humans can specify the conditions of their judgment by, for example, requiring persons to make informed, conscious decisions that are based on sufficient information about the applicable law, the target, the weapon, and the context in which the weapon is deployed.[29] These and more requirements have been specified by different actors in the discourse on autonomous weapons. The targeting process provides opportunities for humans to actually exercise these key elements or components of control in an organised and structured manner.

Conclusion

In this short article I analysed the targeting process, the manner in which humans currently exercise control, and the potential human-machine relationship. I used an example of an existing weapon system with autonomy in its critical functions to see how 'Meaningful Human Control' is understood and implemented as of today. As autonomous weapons do not operate in a vacuum, an illustration of the targeting process is indispensable in order to gain a better understanding of the concept of 'Meaningful Human Control'. The example of the Harpy loitering weapon demonstrates that even though the system selects and attacks a target in phase 5 without any human intervention, it is not truly autonomous in the overall targeting process. However, the analysis should not end at this point. Due to rapid technological advances, autonomous functions and their complexity will change. Although it seems unlikely that anyone would desire a weapon that autonomously executes the entire targeting process without human involvement whatsoever, the possibility of a machine-dominated targeting process must be taken seriously, no matter how unlikely it sounds or how far ahead in time that scenario may be. Therefore, autonomous weapon systems with complex reasoning, machine learning, self-awareness, or a meaningful level of artificial intelligence of sorts should also be considered in light of the targeting process in order to assess how humans can remain in control of these increasingly autonomous weapons.

In addition, one of the effects of increasingly autonomous weapons seems to be further distribution (or redistribution) of tasks. As these tasks become more distributed, individuals tend to become part of an increasingly long chain of human actors and technologies, in which decisions of entities elsewhere in the chain affect the choices or control that others

28 Article 36, Killing by Machine – Key Issues for Understanding Meaningful Human Control, 2015, http://www.article36.org/wp-content/uploads/2013/06/KILLING_BY_MACHINE_6.4.15.pdf.

29 M.C. Horowitz and P. Scharre, Meaningful Human Control in Weapon Systems: A Primer. Working Paper, Center for a New American Security, 2015, https://s3.amazonaws.com/files.cnas.org/documents/Ethical_Autonomy_Working_Paper_031315.pdf, p. 4.

have.[30] For example, technically advanced LAWS may eventually be programmed to decide independently and on-the-spot (i.e. in real time) that a target is a military objective by nature.[31] This means that legal considerations about distinction and military advantage are to be made in the design and validation processes as well. Therefore, besides the targeting, other decision-making processes in which humans make judgments about the use of these weapons also deserve attention in the analysis of 'Meaningful Human Control' (e.g. design process, validation and verification processes, and Article 36 legal review processes).

It is up to humans to discuss these scenarios in depth and decide where to draw the line between what is 'Meaningful Human Control' and what is not. Here, I offered one way of looking at the concept of 'Meaningful Human Control' by focusing on the context of the use of autonomous weapons. I presented an analysis of the targeting process to explain that autonomous weapons do not operate in a vacuum. Moreover, the targeting process provides opportunities for humans to exercise key elements or components of control over increasingly autonomous weapons in an organized and structured manner. Among other decision-making processes, the targeting process is one that should be considered when thinking about increasingly autonomous weapons and the manner through which humans can remain in control of them.

30 M. Noorman, Responsibility Practices and Unmanned Military Technologies, 20 Science and Engineering Ethics Journal 809 (2013) 812.

31 Such autonomous weapons (which would conduct target development and validation in a way that is currently conducted by humans in phase 2 of the targeting process) do not yet exist. Whether this is technologically feasible is, of course, relevant to the debate. However, technological feasibility should not distract us from the question of control. A negative answer does not relieve us from the responsibility to address questions about how we would remain in control of increasingly autonomous weapons.

International Humanitarian Law, Article 36, and Autonomous Weapons Systems

Christopher M. Ford *

Introduction [1]

Article 36 of Additional Protocol I (API) requires States Parties to review weapons[2] in order to ensure their compliance with international law.[3] This obligation is generally accepted to reflect customary international law.[4] International Humanitarian Law (IHL) applies to autonomous weapons systems[5] and imposes requirements on parties to a conflict before and during an attack. This includes requirements that the weapons be reviewed and that

* Lieutenant Colonel, Stockton Center for the Study of International Law at the U.S. Naval War College. The views expressed are the author's own and do not represent the U.S. Department of Defense or any other government entity.

1 This piece is drawn from research conducted as part of a larger research project in the Stockton Center for the Study of International Law at the U.S. Naval War College. Given our unique position as both an independent academic entity and a member of the U.S. Department of Defense, we have had the fortune to receive input and guidance from some of the leading academic, government, technology, and military experts on autonomous technologies. Their assistance has been invaluable and I would like to publicly thank everyone who has provided assistance on this project.

2 According to its text, Article 36 Additional Protocol I applies to weapons, means or methods of warfare. For the purposes of this piece, these terms will be used interchangeably.

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

4 International Committee of the Red Cross (ICRC), A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977, 88 International Review of the Red Cross 864 (2006) ("It seems highly likely that there is a customary rule requiring all states to review new weapons…"); M.N. Schmitt (ed.), Tallinn Manual on the International Law Applicable to Cyber Warfare, Cambridge 2013, Rule 48 ("All States are required to ensure that the cyber means of warfare that they acquire or use comply with the rules of the law of armed conflict that bind the State."); Program on Humanitarian Policy and Conflict Research, Manual on International Law Applicable to Air and Missile Warfare, 2009, Rule 9 ("States are obligated to assess the legality of weapons before fielding them…"); W.H. Boothby, Conflict Law: The Influence of New Weapons Technology, Human Rights, and Emerging Actors, The Hague 2014, p. 169 ("It seems highly likely that there is a customary rule requiring all states to review new weapons…").

5 ICRC, International Humanitarian Law and the Challenges of Contemporary Armed Conflicts 40, Report to the 31st Conference of the Red Cross and Red Crescent, Geneva, Switzerland 31IC/11/5.1.2 2011, p. 36 ("There can be no doubt that IHL applies to new weaponry and to the employment in warfare of new technological developments.").

the attack be discriminate,[6] proportionate,[7] and comply with requirements for precautions in attack.[8] Autonomous weapons, however, can operate in ways that raise questions regarding the application of these requirements and the role of the weapons review process in ensuring compliance.

This chapter will first briefly discuss situations in which autonomous technologies do not implicate concerns in IHL. The next section addresses the requirements of IHL and the ways in which autonomy complicates the application of IHL. The final section considers the role of weapons reviews in ensuring autonomous weapons systems comply with IHL.

Situations not implicating IHL

Autonomy can be employed in ways that do not implicate International Humanitarian Law. These include:

- use in *technologies* that do not implicate International Humanitarian Law. For instance, an unarmed system may autonomously fly from Point A to Point B;
- utilisation in weapon systems in all functions, but the systems are employed in an *environment* that does not implicate aspects of International Humanitarian Law. For example, an autonomous weapons system could be employed in a very limited area (e.g. a remote desert battlefield) or in a very limited fashion (e.g. a weapon is activated for a very short period of time when no civilians are present).

6 Art. 51(4) API.
7 Art. 51(5)(b), Art. 57(2)(a)(iii), and Art. 57(2)(b) API.
8 Art. 57 API.

Situations implicating IHL

Distinction

Distinction requires parties to a conflict to distinguish between lawful targets (combatants, civilians taking direct part in the hostilities, and military objectives) from unlawful targets (civilians, those *hors de combat*, civilian objects, and other protected persons and objects).[9] Autonomous technologies could create a situation in which the distinction determination is made by a computer executing an algorithm, perhaps long after the system was activated. This raises significant questions with regards to distinction:

- When does the system conduct the analysis – when it is activated or when it is about to engage?
- Is the system sufficiently sophisticated to distinguish civilians from combatants?
- To what extent does the operator need to understand the operational parameters of the weapons system?
- How does the environment in which the system is operating affect its ability to conduct a distinction analysis?
- Can an autonomous system identify a civilian who is directly participating in hostilities?[10] If so, can the system identify the period of deployment and redeployment before and after the act of hostilities?

Autonomy also raises questions regarding the targeting of some military objectives.[11] Only military objectives are valid targets. Article 52(2) of Additional Protocol I, which non-parties to the instrument regard as an accurate reflection of customary law, sets forth a two-part test for qualification as a military objective. First, the objects "by their nature, location,

9 Art. 57(2)(a)(1) API (The article requires the attacker to "do everything feasible to verify that the objectives to be attacked are neither civilian nor civilian objects and are not subject to special protection but are military objectives (...)").

10 Y. Sandoz et al. (ed.), Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949, Geneva 1987, para. 1944 (direct participation in hostilities refers to a very broad category of "acts of war which by their nature or purpose are likely to cause actual harm to the personnel and equipment of the enemy armed force.").

11 Art. 52(1) API (civilian objects are defined as all objects which "are not military objectives.").

purpose or use make an effective contribution to military action". Secondly, their "total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage".[12]

Autonomy is unlikely to create issues with regards to objects that amount to military objectives by *nature*[13] or *location*[14] since they are readily identifiable and can be easily programed into an autonomous system. Further, objects by their nature are not susceptible to changes to the operational environment. A tank does not lose its status as a military objective because of a particular event on the battlefield.

Objects that are military by their *purpose* and *use*, however, present difficult challenges for autonomous weapons systems. The purpose of an object speaks to its future use.[15] This determination is based on current intelligence and knowledge of the enemy's tactics, techniques, and procedures. The validity of a target based on its military purpose is thus dynamic. Qualification of objects as valid targets by their use is likewise dynamic. Use indicates the current function of the object, something that can change.[16] When using an autonomous system to attack objects that are valid targets by the purpose and use criteria, measures would have to be taken to ensure they are attacked only during the time that they amount to military objectives.

Proportionality

Even when a target is lawful, the attack itself must comply with the principle of proportionality, which prohibits an "attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which

12 Art. 52(2) API.

13 ICRC Commentary (n 10), para. 2020 ("This category comprises all objects directly used by the armed forces: weapons, equipment, transports, fortifications, depots, buildings occupied by armed forces, staff headquarters, communications centres etc."); United States Department of Defense Manual, Law of War, 2015, p. 209 ("'Nature' refers to the type of object and may be understood to refer to objects that are *per se* military objectives.").

14 ICRC Commentary (n 10), para. 2021 ("Clearly, there are objects which by their nature have nonmilitary function but which, by virtue of their location, make an effective contribution to military action."); DoD Manual (n 13), p. 209 ("The location of an object may provide an effective contribution to military action.").

15 ICRC Commentary (n 10), para. 2022 ("The criterion of 'purpose' is concerned with the intended future use of an object, while that of 'use' is concerned with its present function.").

16 DoD Manual (n 13), p. 209 ("'Use' refers to the object's present function.").

would be excessive in relation to the concrete and direct military advantage anticipated".[17] An autonomous system may have the capability to deploy for long periods without human operator interaction. This generates significant International Humanitarian Law questions:

- Who calculates military advantage, the operator or the machine?
- If the system is conducting the analysis, can it receive or perceive changes to the anticipated military advantage and collateral damage?
- Can the military advantage assessment be reduced to a mathematical calculation?

Military advantage is calculated with a view to "the circumstances ruling at the time".[18] Systems cannot be preprogramed with all necessary information to make this calculation, particularly where systems are deployed for long periods of time. Some have questioned whether military advantage could be reduced to a mathematical formula that could be programed into an autonomous weapons system.[19] It would appear that it can for a discrete period of time. Militaries often conduct a value analysis of a particular target. For instance, in the establishment of a prioritised list of targets, targets are being valued against one another.[20] Military advantage requires a combatant to assign a value of military importance for a given target. These values might change as the military advantage shifts, but there is no reason to think that values cannot be assigned in 'good faith' to military targets *ex ante*.[21]

The ability to update military advantage could be fulfilled through a number of possible mechanisms:

- *sophisticated system*: deploying a system that is sufficiently advanced so as to be able to perceive and understand changes to the military advantage;
- *updates*: updating the system as needed, which might be constantly, or something less, depending on the nature of the battle;

17 Art. 57(2)(a)(iii) API; see also Art. 51(5)(b) API ("an attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."); DoD Manual (n 13), p. 241 ("Combatants must refrain from attacks in which the expected loss of life or injury to civilians, and damage to civilian objects incidental to the attack, would be excessive in relation to the concrete and direct military advantage expected to be gained.").

18 Art. 52(2) API.

19 W.H. Boothby, The Law of Targeting, Oxford 2012, p. 96.

20 See e.g. United States Department of Defense Joint Publication 3-60, Joint Targeting, 13 April 2007, II-8 ("A [joint target list] is a prioritized list of targets (…)").

21 ICRC Commentary (n 10), para. 2208 ("the [military advantage] interpretation must above all be a question of common sense and good faith for military commanders.").

- *human*: retaining a human on or in the loop;
- *restricted use*: deploying the system for a sufficiently short period of time so that the military advantage will not change during the deployment.

Precautions in attack

An attacker must further comply with Article 57 of Additional Protocol I, precautions in attack, which requires "constant care" to be taken to "spare the civilian population, civilians and civilian objects".[22] This article is widely considered to reflect customary international law.[23] The Protocol does not define "constant care", but this phrase suggests something more than a one-time obligation.[24] There are several mechanisms that could be applied to autonomous weapons systems to ensure compliance with this requirement:

- constant care could be exercised throughout the operation based on programmed instructions. For example, an autonomous system could be programmed to only target enemy tanks when no civilians are present;
- constant care could also be exercised by constantly monitoring the system, or using an 'ethical governor' in the programming of the system.[25]

Article 57 further requires that those or who "plan or decide upon an attack (...) do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects and are not subject to special protections but are military objectives (...)".[26] A prolonged deployment of an autonomous system begs the question of when the requirement to take such feasible verification measures applies – when the system is activated, when the

22 Art. 57(1) API.

23 J.M. Henckaerts and L. Doswald-Beck (ed.), Customary International Humanitarian Law, Cambridge 2005, rule 15; see also Tallinn Manual (n 4), p. 165 (this rule "is considered customary in both international armed conflict and non-international armed conflict.").

24 See Tallinn Manual (n 4), p. 166 ("Use of the word 'constant' denotes that the duty to take care to protect civilians and civilian objects is of a continuing nature throughout all cyber operations.").

25 See e.g. R. Arkin et al., An Ethical Governor for Constraining Lethal Action in an Autonomous System, Technical Report GIT-GVU-09-02, http://www.cc.gatech.edu/ai/robot-lab/online-publications/GIT-GVU-09-02.pdf.

26 Art. 22 of the 1907 Hague Regulations Respecting the Laws and Customs of War on Land; see also Art. 57(2)(a)(i) API.

system is about to engage, or throughout the targeting process? The better view is that the obligation is continuous in nature and runs beginning with the programing of the software through the length of the engagement.[27]

Weapons reviews of autonomous weapons systems

Article 35 of Additional Protocol I reaffirms the longstanding proposition that the methods and means of warfare are not unlimited.[28] This concept is operationalised by Article 36 of Additional Protocol I which imposes on States Parties an obligation to ensure they do not use unlawful weapons. The mechanism for this obligation is the weapons review process. While the Additional Protocol does not mandate the form of the weapons review, it is widely accepted that a review should consider both the weapon itself and the anticipated "normal or expected use" of the weapon.[29]

Weapons reviews of autonomous systems are complicated because of three inter-related factors: (1) the complexity of the underlying technology; (2) the potential for unpredictable systems; and (3) the possibility of autonomous systems that can learn and adapt. While not unique to autonomous weapons, the implicit technological sophistication of autonomous systems demands increasingly sophisticated means of testing the systems. This highlights several issues in the testing process, including:

- How are technical results from tests translated into a narrative that a legal advisor can understand?
- How are tests designed for new technologies?
- How can tests replicate the normal and expected use of the weapon?

27 M. Bothe et al., New Rules for Victims of Armed Conflicts, The Hague 1982, p. 363; see also Art. 57(2)(b) API (an attack must be "cancelled or suspended if it becomes apparent that the objective is not a military one or is subject to special protection or that the attack may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated.").

28 Art. 35 API.

29 ICRC Commentary (n 10), para. 1469; see also ICRC (n 4), p. 938 ("A weapon or means of warfare cannot be assessed in isolation from the method of warfare by which it is to be used. It follows that the legality of a weapon does not depend solely on its design or intended purpose, but also on the manner in which it is expected to be used on the battlefield.").

- How can a weapons review be conducted on a weapon that is so complex that it is physically impossible to test all lines of computer code?
- How does testing account for design and manufacturing errors?[30]

International humanitarian law prohibits two broad categories of weapons: those that cause superfluous injury or unnecessary suffering,[31] and those that are inherently indiscriminate, including weapons that cannot be aimed or whose effects cannot be controlled.[32] Both categories of prohibitions reflect customary international law.[33] Autonomous weapons systems are unlikely to run afoul of these provisions. There is nothing inherent in an autonomous weapons system that raises unique issues with regards to the prohibition against superfluous injury or unnecessary suffering, and a weapon will only be indiscriminate when the weapon is incapable of being used discriminately. Weapons reviews should, of course, consider these prohibitions, though they are unlikely to be violated by the autonomous aspect of the weapon.

Reviewing the lawfulness of an autonomous weapon in the context of its expected and normal use, however, raises several questions. Consider, for instance, questions related to distinction that might arise in an autonomous weapons review:

- How well can the system distinguish between civilian and combatant and between civilian object and military object? How is this quantified?
- How do changes in the *physical* environment (e.g. atmospheric conditions, time of day, and weather) affect the ability of the system to distinguish?
- How do changes in the *operational* environment (e.g. the persons and man-made structures which are physically present) affect the ability of the system to distinguish?

30 See generally A. Backstrom and I. Henderson, New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews, 94 International Review of the Red Cross 886 (2012) (providing an analysis of the technical issues which arise when conducting weapons reviews for highly sophisticated weapons).

31 Art. 35(2) API.

32 Art. 51(4)(b) and Art. 51(4)(c) API.

33 J.M. Henckaerts and L. Doswald-Beck (ed.), Customary International Humanitarian Law, Cambridge 2005, rules 70 and 71; see also W.H. Boothby, Weapons and the Law of Armed Conflict, 2nd ed., Oxford 2016, pp. 46-73.

Significant questions are also raised when considering proportionality and precautions in attack:

- To what extent is the system making proportionality calculations (as opposed to calculations being pre-programmed or made by a human operator)?
- Is the system sufficiently sophisticated and reliable such that it can make a collateral damage estimate?
- How does the system account for changes to the military advantage?
- Can the attack launched by the autonomous weapons system be cancelled or suspended; and if so, who makes the determination, the system or the operator?[34]
- Can the autonomous weapons system perceive when a warning is required under Article 57(2)(c)?[35]
- Does the system have the ability to select among several similar military objectives so as to cause the least danger to civilian lives and civilian property?[36]

The weapons review will require careful consideration of the technology used in the system, the nature of expected use, and the environment in which the system is expected to be used. Plainly, some systems may be lawful in particular circumstances or in certain environments. In all reviews, there are certain best practices which should be considered. This includes:

- the weapons review should either be a multi-disciplinary process or include attorneys who have the technical expertise to understand the nature and results of the testing process;
- reviews should delineate the circumstances of use for which the weapon was approved;
- the review should provide a clear delineation of human and system responsibilities. Who will do what in a given circumstance?
- optimally, the review should occur at three points in time. First, when the proposal is made to transition a weapon from research to development. Second, before the weapon is fielded.[37] Finally, weapons should be re-reviewed periodically based upon feedback on how the weapon is functioning. This would necessitate the establishment of a clear feedback loop which provides information from the developer to the reviewer to the user, and back again.

34 Art. 57(2)(b) API.
35 Art. 57(2)(c) API.
36 Art. 57(3) API.
37 This two-step review is the process adopted by the United States Department of Defense Directive on Autonomous Weapons; United States Department of Defense, Directive 3000.09, Autonomy in Weapon Systems, 21 November 2012.

Lethal Autonomous Weapons Systems: Proliferation, Disengagement, and Disempowerment

Jai Galliott*

Introduction

In this chapter, I will speak to the irrationality of an outright ban on lethal autonomous weapons systems and, in so doing, wish to make three points, the first pertaining to the value that stands to be derived from autonomous systems. One of the most frequently touted benefits of exploiting robotics is that they mitigate the human cost of war, assessed physically, psychologically, or otherwise, with the level of mitigation largely dependent on the degree of separation between the operator and their system. In our discussions, it must be remembered that many object to the use of robots with humans directly 'in the loop' on the basis that the psychological impact on operators is just as devastating as physically fighting war from an in-theatre location. The employment of increasingly autonomous systems could counter this problem. More generally, such systems seem particularly unproblematic if used in clearly demarcated conflict zones such as the Korean DMZ or in circumstances in which enemy targets are located remotely and clearly differentiated from non-combatants, civilian infrastructure and the like.[1] While such situations are indeed rare in the age of asymmetric conflict, in which combatants and non-combatants commingle, my personal belief is that it would be unwise to prevent states from using autonomous systems where and when such circumstances arise, especially in the broader context of enhancing overall military effectiveness and efficiency in the pursuit of just outcomes, and given that they yield environmental benefits as well, in that they reduce fuel consumption, pollution and otherwise stand to limit the damaging human footprint of more conventional war.

More importantly, and as my second point, it makes little sense to propose a ban on lethal autonomous weapons that "select and engage targets without human intervention", as some propose, because when you think about it critically, no robot can really kill without human intervention. Yes, robots are already capable of killing people using sophisticated mechanisms and procedures that *resemble* those used by humans, meaning that humans

* The University of New South Wales, Australia.

1 J. Galliot, Military Robots: Mapping the Moral Landscape, Surrey 2015, p. 235.

do not necessarily need to oversee a lethal system while it is in use. But that does not mean that there is no human in the loop. It would seem that Alan Turing was largely correct in predicting "that by the end of the century the use of words and general educated opinion will have altered so much that we will be able to speak of machines thinking without expecting to be contradicted".[2] That is to say that while we can model the brain, human learning and decision-making to the point that these systems are capable of mimicking human problem-solving ability and finding computational solutions to killing people, humans are very much involved in this process. Indeed, it would be preposterous to overlook the role of programmers, engineers and others involved in building and maintaining these autonomous systems. And even if we did, what of the commander or politician who assesses the context in which an autonomous system is initially deployed? No matter how 'intelligent' they may become or how well they 'learn', machines will always lack genuine human-level decision-making capability and consciousness, therefore meaning that humans should be the focus of our attention. It is for this reason that the development of autonomous weapons merits caution, rather than a blanket prohibition on their use.

However, my main fear and third substantive point (developed below), put in the language of recent expert meetings, is that attempts to define 'autonomy' and 'meaningful human control' in terms that are conducive to a complete prohibition on lethal autonomous weapons systems might distract or even prevent nations from taking what could be more effective action to regulate rather than ban the relevant technology. I submit to you that we already have lethal autonomous weapons systems of the kind that many are inclined to ban and that the biggest risk to international security is not the proliferation of these systems to rogue states and non-state actors, but rather their proliferation within established western armed forces, primarily because of their potential to divert investment away from soldiers, as the core of any armed force, in a way that impacts core combat proficiencies, limits force readiness and/or degrades conventional warfighting capability that may need to be relied upon if autonomous systems fail or cannot be deployed in a particular conflict scenario. As we all know, there are limits associated with all robotic solutions to military problems, most related to technical limitations, hacking, interoperability, cultural acceptance issues and others, any one of which holds the potential to convert or revert technological conflict into more conventional warfare. Military and political leaders must therefore determine how vulnerable they would be if they substantially invest in robotics only to find that the technology fails or is otherwise made redundant.

2 A. Turing, Computing Machinery and Intelligence, 59 Mind 433 (1950) 435.

The limits of robotic solutions

It is important for nations considering employing lethal autonomous weapons systems to understand that while machines can do some things that humans cannot, there are a number of respects in which they have their limits and can actually hinder military operations. One of the key concerns with robotic systems is that their sensors may be insufficiently robust to achieve mission success in all operating environments. Even the best robotic systems with high quality sensors can be rendered largely ineffective against simple countermeasures like smoke, camouflage, and decoys, or by naturally occurring features such as bad weather and infrared sensor-degrading foliage.[3] As a result of operations in Afghanistan, Iran, Iraq and Pakistan, among others, intelligent state and non-state groups have also developed standard operating procedures for defeating or mitigating the effectiveness of emerging technologies, regularly exploiting mountainous regions, cave systems and sprawling bunkers, allowing them to avoid detection and capture.[4]

Even if system developers overcome sensor limitations, it is imperative that robotic systems are capable of operating at extended range, and this calls for a secure network link and a reliable way to process the signals it transfers. Untethered robotic systems require persistent, high-bandwidth communication lines to allow for the upload and download of data, and these can be difficult to maintain in peacetime, let alone in mountainous deserts or the congested electromagnetic spectrum of urban military environments. Latency – the time it takes to process a signal into the correct system output (say, firing a missile) – has also proven to be an issue and caused some military robotics acquisition programs to be abandoned in their entirety.[5]

Although the ideal method of operation for many forces involves a transition to increased autonomy, this is only a partial solution, since even highly autonomous systems will require some kind of persistent connection to enable the system access bulk data relevant to targeting, if not to allow human military and political leaders to maintain situational awareness and abort authority. Communications researchers are investigating novel solutions to these problems by investigating the full use of the electromagnetic

3 S. Biddle, Afghanistan and the Future of Warfare: Implications for Army and Defense Policy, Strategic Studies Institute, 2002, p. 32.

4 S. Kay, Global Security in the Twenty-First Century: The Quest for Power and the Search for Peace, Lanham 2015, p. 202; R. van den Berg, The 21st Century Battlespace: The Danger of Technological Ethnocentrism, 10 Canadian Military Journal 10 (2010) 14.

5 B. Crispin, What Killed the Robot Soldier, 10 November 2008, http://www.strangehorizons.com/2008/20081110/crispin-a.shtml.

spectrum and improving bandwidth, but physical limits (speed-of-light, horizon line-of-sight, etc.) will always result in signal delays and any increase in bandwidth will be directly proportional to the ease with which that signal can be hijacked and blocked.[6] Any such attack on data communication systems could be extremely detrimental given the nature of their interdependence. Even the temporary interruption or disablement of a robot or its host system by a directed-energy laser or electromagnetic pulse weapon could reduce trust in the system and slow information dissemination, with grave consequences for information-dependent operations.[7] In the worst case scenario, a major breach of a vulnerable system could cause a loss of connectivity between the individual elements of a military force or indeed whole armies.

Needless to say, connectivity is critical to interoperability when and where there are fewer humans in the loop, and the degree of interoperability that most modern militaries seek to attain can only be achieved by operating on a tightly integrated battle network in order permit the free flow of information. Nevertheless, the encryption of communications lines is the only method to ensure their integrity, but this is a double-edged sword in the sense that doing so compromises range and the speed with which systems can enact tactical commands.[8] The robot security requirement is also in direct contest with systems architecture requirements, namely that autonomous systems be capable of operating in synergy with a wide range of capabilities from a vast multi-national ecosystem of hardware and software suppliers because, while this can be achieved by purchasing systems with common components and software programs or by developing systems to common (or possibly even 'open') standards, such measures increase the risk that a cyberattack that is successful upon one system could bring down all systems of a particular nation and those of allied nations that have based their systems on the same standard, with potentially disastrous consequences for the future of international coalitions. The integration of robotics into one's force structure also presents a further, more general, interoperability dilemma for some Asian, European, and Oceanic nations that may in the future require the support of developing nations likely to have less advanced capabilities and that lack the research and development or acquisition funds to purchase and operate advanced robotics technologies. Therefore, if a problem which requires non-super power coalition fighting should arise, robotics-enabled states may be less capable of entering combat interoperability with other nations' military equipment, as demonstrated by the U.S. when it discovered that the com-

6 P. Hew, The Generation of Situational Awareness within Autonomous Systems – A Near to Mid Term Study – Issues, Defence Science and Technology Organisation, Canberra 2010, p. 6.
7 Kay (n 4), p. 207.
8 Crispin (n 5).

munications systems within its fighter aircraft where incompatible with those of its North Atlantic Treaty Organization (NATO) allies.[9] This means that autonomous robotics-wielding states will not always be able to fully utilise their technological advantage in war.

In other cases, it may be prudent to limit the use of autonomous systems on cultural acceptance grounds. On the internal front, a massive cultural shift will be required within most military forces if they are to exploit any form of robotics with success. This is particularly true of land forces, which are typically the least technologically advanced arm of tri-service defence organisations and employ thousands of personnel who may have spent their whole lives – social, physical and psychological – preparing or maintaining readiness to defend their nation by more traditional means and now stand to be displaced by technology. Until acceptance levels amongst soldiers and their commanders match doctrine and standard operating procedure, it is likely that technologically enhanced operations will fail or achieve limited success. On the external front, socio-cultural sensitivities are important. As the U.S. learned in Iraq and Afghanistan, failure to secure the oft-cited 'hearts and minds' of locals can lead to the unwillingness of said locals to cooperate with international forces and transitional governments, such as the one initially implemented in Iraq.[10] Going forward, other nations may find it problematic to secure the emotional and intellectual support of certain foreign populaces if their forces include autonomous systems, for it is clearly imaginable that, for those civilians in the danger zone, this will be reminiscent of previous US efforts to send robots to fight their wars and kill their enemies in a way that did not sit well with those on the ground. Enemy combatants might take it to be demonstrative of the fact that their counterparts are not willing to settle their disputes in the customary, and somewhat more chivalric, face-to-face manner. It is known that the Taliban, for instance, accustomed to fighting in close proximity to their Middle Eastern enemies, see the use of these distance weapons as extremely cowardly.[11]

This may also hold true in relation to civilians, for many will have already encountered the powerfully alienating hum of Predator drones flying overhead, and seen the side effects of their use, without ever seeing the faces of those operators behind them.[12] There is also the

9 Kay (n 4), p. 201.

10 J. Cummins, Why Some Wars Never End: The Stories of the Longest Conflicts in History, Beverly, MA 2010; A.C.G.M. Robben, Losing Hearts and Minds in the War on Terrorism, in A.C.G.M. Robben (ed.), Iraq at a Distance: What Anthropologists Can Teach Us About the War, Philadelphia 2010, p. 106-132.

11 L. van Wifferen, Alienation from the Battlefield: Ethical Considerations Concerning Remote Controlled Military Robotics, Utrecht 2011.

12 Center for Civilians in Conflict, The Civilian Impact of Drones: Unexamined Costs, Unanswered Questions, New York 2012, p. 23.

risk that those who have not had such experiences will be guided by unbalanced media reporting of drone strikes of a kind that autonomous systems will not replicate. Moreover, for every autonomous system that supplements a soldier, there will be one less human on the ground to take responsibility for any collateral damage and carry forward values like honour, honesty, duty, respect, integrity, and personal courage, which are traditionally attributed to warriors who are considered to be representatives of their respective states.

Beyond highlighting that robotic solutions have their limits, the overall concern is that if a number of armed forces were to become an advanced lethal autonomous weapons force, the result could well be militaries that are collectively dominant against an all too particular type of enemy force in equally select theatres of war, but irrelevant to the emergence of non-state groups and modern-day warrior societies that will likely need to be tackled with human soldiers rather than technology. Indeed, it would be all too easy for conflicts to be initiated with the view that they can be won by an autonomous or roboticised force with a reduced number of humans placed in harm's way, only to find at a later stage that victory – or the abortion of the operation – can only be achieved by committing a much greater number of soldiers or manned systems to the zone of hostile operations. This could occur for any of the reasons noted above, because the system(s) malfunctioned, or because changing circumstances result in the operation being beyond the reach or capability of the system(s) in question, or because the technology-focused operation was ill-conceived or impractical to begin with.[13] History is replete with examples demonstrating that investment in technology over human beings will not necessarily equal decisive victory. For example, many consider the Vietnam War to be proof that technologically induced complacency and dependency can lead to disapproval on the home front and, ultimately, failure. Critics wrote of "blind technological fanaticism, hubris, and overconfidence as the United States attempted to fight a remote, antiseptic war".[14] The strategy in this conflict was initially one of airpower dominance, but America underestimated just how ineffective its technologically sophisticated fighter jets would be over a land covered in dense jungle foliage, and the extent to which such technology would antagonise the civilian population, without whose political support victory was not possible. And, as already mentioned, Iraq and Afghanistan are perhaps more recent examples of the resounding effects of technological complacency and dependency. In Lawrence J. Korb's words, the idea that war could be won with the effort of "three guys and a satellite phone", proved to be quite dangerous.[15] In

13 R. Sparrow, Building a Better Warbot: Ethical Issues in the Design of Unmanned Systems for Military Applications, 15 Science and Engineering Ethics 169 (2009) 173.

14 K.P. Werrell, Did USAF Technology Fail in Vietnam?, 12 Airpower Journal 87 (1998) 87.

15 Quoted in P.W. Singer, Wired for War: The Robotics Revolution and Conflict in the 21st Century, New York 2009, p. 316.

all of these conflicts, infantrymen were paramount to the cessation of conflict. It needs to be recognised that even when confronting an enemy with technological forces, there is a possibility that the distance-oriented autonomous warfare that robotics and other technologies enable could collapse into more traditional warfare where well-trained soldiers are required at an additional cost to any investment in technology and, as argued in the next section, provisions for said personnel might not be available in the required generation cycle if a nation fails to prepare for such possibilities and over-relies on – or over invests in – lethal autonomous weapons systems.

Dependence, deskilling, and degradation: compounding the limitations of autonomous systems

In the above section it has been argued that when applying technology to military problems, nations must understand the limitations of technology and determine how vulnerable they would be if they substantially invest in robotics only to find that the technology fails or is otherwise made redundant in a particular conflict scenario. In order to make this determination, and prior to considering how any vulnerabilities might be offset, those contemplating the use of lethal autonomous weapons systems must consider the extent to which artificial intelligence might inadvertently come to substitute fundamental skills in training and, at this juncture, ask whether soldiers would have the basic warfighting skills to operate without artificial intelligence (AI) and autonomous systems technologies in the case of their failure. Commanders and education units currently face the challenge of preparing military personnel for real world operations without knowing exactly what the next conflict will look like, but they must nevertheless be capable of handling virtually any situation and, given the increasingly unpredictable nature of world politics, be prepared to do so at short notice. Unfortunately, as this section will show, technology can serve as a temporary and often inadequate capability support mechanism while proper training is sacrificed and core combat skills atrophy.

With regard to the skills maintenance problem, we can learn valuable lessons from the United States, which has become increasingly reliant upon these robotic systems and, in some instances, found itself unable to reach mission success on the basis of their core training and human ingenuity alone. This technological dependence creates a risk that military personnel may not be able to overcome themselves, especially when coupled with the inadequacies that inevitably arise with the acquisition of more autonomous technologies and which often take significant time to resolve. The ease of access to technology is certainly a contributing factor. Land navigation is particularly at risk in this sense because

the GPS units available to most soldiers, and which are at the heart of remotely operated robotics, enable those operating in the technological battlespace to ignore or otherwise fail to maintain basic skills pertaining to map reading, terrain studies, or orientation.

In relation to this matter, Raymond Millen has written that "the general attitude is that these devices have made much training unnecessary".[16] While all of this is true, the problem is, in some respects, more systemic in nature. A recent internal study of a U.S. Army regiment's performance in manoeuvres at the Joint Multinational Readiness Center, a training ground in Germany, found that the regiment's basic combat skills had extensively deteriorated as a result of the unit having only recently operated out of hardened command posts with strong digital connectivity.[17] The report states that units "rely heavily on digital systems and seem at a loss when they must revert to analogue or manual systems for long stretches of time, resulting in a total loss of situations awareness", and that commanders were "tethered to the command post" rather than executing battlefield circulation.[18] More generally, the report was critical of senior leaders' understanding of their command role, right through to privates' understanding of basic duty functions and even basic field hygiene requirements. It is this phenomenon of dependency and complacency that other nations need to circumvent as the use of lethal autonomous weapons systems and other emerging technologies steadily increases over time.

In many respects, the use of highly autonomous robotics is likely to significantly exacerbate the general skills maintenance problem presented by technology, because the skills required to operate, oversee, or deploy these systems significantly diverge from those required to operate semi-autonomous systems and less complex tools such as GPS, blue force trackers, and so on. That is to say that while some of the tasks that personnel accomplish using autonomous systems may be very similar to those they perform when using or being supported by the non-autonomous equivalent – meaning that there is little or no difference in the skills or training required to maintain proficiency in both forms of operations – there will be many more cases where the tasks are very different and a combination of skills, or a completely different skillset, is required. Acquiring these non-transferable skills and preparing for autonomous robot-enhanced operations can take significant time away from maintaining and enhancing one's traditional soldiering skills and, in the case of hurriedly adopting technology and/or fighting in technologically-enhanced combat, training

16 R. Millen, The Art of Land Navigation GPS Has Not Made Planning Obsolete, 90 Infantry Magazine 36 (2000) 36.
17 J. Crafton, Collection Report: Decisive Action Training Environment 2nd Cavalry Regiment (Stryker), Joint Multinational Readiness Center, Bavaria 2012.
18 Ibid., p. 5.

opportunities may be reduced for months or even years, and the average infantryman, as well as personnel from units with special skills or equipment, will be unable to practice some critical combat skills. We have already seen this with regard to rated pilots from manned airframes who have been selected to 'cross train' for unmanned systems service. These pilots typically hold a significant amount of flying experience. They serve on a drone assignment for three to four years.[19] Over the years, these experienced pilots may come to be deprived of any real-world flying experience or training and, if they are for some reason forced to return to traditional air combat in closer proximity to the enemy, they may be susceptible to being harmed more easily and be less effective in fulfilling their obligations to the military. Over the long term, as autonomy improves and more robots come to operate without the intervention of a soldier, it is also likely that there will be a skills transference from traditionally trained military members who hold rarely-practiced combat skills developed primarily in the early years of their careers and applied to robotics in their later careers, to a new era of combatants that will likely see the development of the traditional soldiering skills developed by the generation before them as a burden rather than skills that enhance the employment of technologies such as robotics.

The deskilling could be further compounded if it extends to the soldier's character. The loss of the traditional soldiering or piloting context might mean, for instance, that we also lose the opportunity to cultivate key moral virtues that contribute to a military member's character and the warrior ethos, cardinal examples of which include wisdom, honesty, courage, integrity, loyalty, patience, and moderation. This is concerning because, for the many reasons already outlined, it is the individual soldier who will, for the time being, be the fall-back (if not active) defence mechanism for his or her nation, and who must be most unconditional in exercising moral restraint and adhering to rules of war. It is the soldiers, more than any other war-making agential group, who have the most influence on a war's outcome, as well as the ability to introduce a moral component to military decisions. In Michael Ignatieff's words, "the decisive restraint on inhuman practice on the battlefield lies within the warrior himself – in his conception of what is honourable or dishonourable for a man to do with weapons".[20] Ironically, soldiers are the primary agents of physical violence, compassion, and moral arbitration in war. The specific concern is that, if this remains the case for some time, and soldiers are morally deskilled, it may lower their ability or willing-

19 W. Chappelle et al., Important and Critical Psychological Attributes of USAF MQ-1 Predator and MQ-9 Reaper Pilots According to Subject Matter Experts, Dayton, OH 2011, pp. 5-7.

20 M. Ignatieff, The Warrior's Honour: Ethnic War and the Modern Conscience, London 1998, p. 118.

ness to adhere to the rules of war and generally lead to unethical decision-making and/or lower barriers to war, thus endangering the moral conduct of warfare itself. It is not difficult to imagine how this could occur.

Shannon Vallor asks us to consider that the skillset for supervising and vetoing the targeting decisions of a semi-autonomous robot will be much narrower than the skillset that would be required if a human being were out in the field targeting the person themselves.[21] One reason for this is that the time the human 'on the loop' has to make moral decisions will in some cases be reduced and, in other cases, will lack information. Some will respond that soldiers in the field often have to make snap decisions, often without all of the available information. However, this misses the point: if the moral skills in the use of lethal force are supplemented by digital systems or not cultivated at the level where force is applied, where and how will they be cultivated and of what value will these skills be in the case of forces that have become dependent on technology? It must be acknowledged that the soldiers of today and tomorrow will need moral as well as technical and core combat skills, and that cultivation of these skills becomes more difficult as each piece of mediating technology enters service and the abilities of autonomous systems are enhanced.

At a workforce analysis level, it also needs to be recognised that the mere perception of there being any form of deskilling amongst today's soldiers will exacerbate existing staffing issues because, even without a spotlight on the concerns related to the impact of robotics, military forces struggle to attract and retain the appropriate number of personnel. In the Australian Defence Force (ADF), for example, there has been a reasonably high level of exit intention. In 2007, it was reported that between 30 and 39 percent of military personnel were "actively looking at leaving the service".[22] More recently, the ADF has confirmed that there has been an increase in the number of people leaving the military, saying that in "2005, 2006 we were talking about a recruitment and retention crisis and seven years later, we are back into a period where our separation rates are increasing".[23] Currently, the ADF utilises targeted retention bonuses, allowances, and initiatives to retain experienced military personnel, but this is only a short to medium term solution and does nothing to alleviate the underlying problem. While there is limited research on workplace issues within the ADF or the military forces of other liberal democratic

21 S. Vallor, The Future of Military Virtue: Autonomous Systems and the Moral Deskilling of the Military, in K. Podins et al. (ed.), 5th International Conference on Cyber Conflict, Tallinn 2013, pp. 471-486.

22 K. Townsend and M. Charles, Jarhead and Deskilling in the Military: Potential Implications for the Australian Labour Market, 34 Australian Bulletin of Labour 64 (2008) 73.

23 ABC News, Number of People Leaving the ADF on the Rise, 16 March 2012, http://www.abc.net.au/news/2012-03-16/adf-confirms-military-exodus/3893826.

nations, American research reveals that skills development is one of the top two reasons why young men join the military.[24] It is therefore easy to understand the frustrations of service members if one joins the military in order to attain certain skills and trains extensively with a view toward developing the acquired skills, only to learn that these original skills will need to be supplemented by an entirely new skillset to work alongside autonomous systems. As a consequence, re-enlistment and retention becomes a problem for some personnel because, for them, there are few 'higher-value' jobs than serving one's nation through rendering active military service of the traditional kind to which they have become accustomed. Any assurance of finding a more meaningful role within the military is likely to be of little consolation. Of course, some military personnel will adapt well and accept positions working alongside robotics, and these are the types required for future military forces. Others will unequivocally reject such roles (consider career pilots who have been pushed to extremes through their training and service, and who value the human element in warfare). After all, a life alongside autonomous robotics involves an entirely different and, for some, potentially less fulfilling form of military service.

Conclusion

All of the above means that our ability to build and maintain true experience of the art of war faces barriers as traditional soldiering functions are handed over to machines and artificial intelligence, so much so that when coupled with the technological limitations of robotics, nations would be wise to add a new dictum to their strategic theorem: the risk of employing technology is that doing so may jeopardise the human element essential to the effectiveness of military forces (especially land forces), and ultimately lead to the demise of the country and people that it serves. The argument here is not that we need to be old-fashioned or that we need to protect ourselves by being big, strong, and as tough as nails, but merely that we must be sceptical of placing our trust in software and machines when the human mind is the most powerful weapon of any military force. Nor is it that we should depend on the mind and brute human force instead of lethal autonomous weapons technologies and technology-orientated tactics – that would indeed be a horrible method of solving military problems, likely to lead to human wave attacks or something akin to the kamikaze. Rather, the argument is that while most skills soldiers hold today will decline in utility over time, there is some value in the 'old way' of soldiering that is inherently necessary in all aspects of combat and which will carry over into wars of the future. To be clear, the classical arts of soldiering, sailing and flying are not completely abandoned in even

24 Townsend and Charles (n 23), p. 74.

the most advanced of today's technological forces, and an increase in the number of lethal autonomous weapons systems will not change this, but more needs to be done to ensure that they are regularly used and maintained as the general skill set of military personnel evolves alongside the changing character of war. I would suggest that further mapping of the 'critical functions' involved in lethal autonomous weapons systems deployment would be most useful in this endeavour and could serve to identify areas that hold potential to decrease conventional combat capability and degrade force strength and readiness. This is not to say that I advocate regulating these functions at an international level. In my view, this is best to level to national policy given the variance likely to form as a result of differences in technology and the way in which different state actors employ it.

Autonomous Weapon Systems and International Law: Consequences for the Future of International Peace and Security

Denise Garcia*

Introduction

The development of lethal artificial intelligence will pose fundamental problems for the stability of the international system and will alter regional security dynamics. States should adopt preventive security governance frameworks, based on the precautionary principle of international law,[1] and on previous cases where prevention brought stability to all countries. Leading scientists working on artificial intelligence have argued that the militarisation and use of lethal artificial intelligence would be a highly destabilising development, as such weapons represent the third major transformation in the nature of war (the first and second transformations being gunpowder and nuclear weapons, respectively).[2] They caution that the deployment of such weapons will alter the future of peace and security for the worse, and they advise that a lethal autonomous arms race could be prevented by a ban on offensive autonomous weapons. It is an important signal when scientists and industry leaders call for a halt in the development of their technology based on what they fear could go wrong.[3]

My analysis is the result of the examination of 22 existing treaties that acted under a 'preventive framework' to establish new regimes of prohibition or control of weapon systems that had been deemed to be destabilising. These treaties achieved one or all of three goals: prevented further militarisation, made weaponisation unlawful, and stopped proliferation with cooperative frameworks of transparency and common rules.[4]

* Associate Professor of Political Science and International Affairs, Northeastern University.

1 G. Marceau, The precautionary principle under WTO law, in: Precaution from Rio to Johannesburg: Proceedings of Geneva Environmental Network Roundtable, Geneva 2002; D. Garcia, Future arms, technologies, and international law: Preventive security governance, 1 European Journal of International Security 94 (2016); D. Freestone and E. Hey (ed.), The Precautionary Principle and International Law: The Challenge of Implementation, The Hague 1996.

2 See Future of Life Institute, http://futureoflife.org/open-letter-autonomous-weapons/.

3 N. Sharkey, Saying ‚No!' to Lethal Autonomous Targeting, 9 Journal of Military Ethics 369 (2010); A. Krishnan, Killer robots: legality and ethicality of autonomous weapons, Farnham 2009; M. Bieri and M. Dickow, Lethal Autonomous Weapons Systems: Future Challenges, CSS Analyses in Security Policy, No. 164 (2014) p. 1.

4 Please contact the author for her database of the treaties mentioned.

As a result of my findings, it is clear that there is a significant emerging norm in regards to all weapon systems: the utilisation of disarmament and arms regulations as a tool and mechanism to protect civilians (civilian protection is the impetus to this new norm).[5] The development of lethal autonomous weapons systems would severely jeopardise this emerging norm. Additionally, in this article, I show under what conditions lethal autonomous weapons systems will be disruptive for peace and security, and show alternative governance structures based on international law with robust precautionary frameworks.

In its 70 years of existence, the United Nations' existence has led to the construction of an edifice of peace and security that shapes and defines international relations in the 21st century.[6] The development of lethal autonomous weapons threatens to undermine the legal and political architecture of peace and security built during the life of the organisation. This edifice of peace and security is built on three pillars, all of which are the result of the intense codification of international law through treaties and the juridification of world politics:[7] the first pillar is the regulation of war and the prohibition of the use of force, with the accompanying global norm of peaceful settlement of disputes between states and increasingly other actors.[8] The second pillar is composed of the dense network of global norms that constrain and guide the behaviour of states.[9] This includes an extensive array of treaties that form numerous governance regimes and comprise an elaborate framework of transparency, confidence building, and security mechanisms put in place to maintain peace and security. The third pillar is built on a structure that fosters cooperation in cultural, economic, social, and environmental matters that affect all of humanity, and tackle problems that need to be solved collectively.

5 L. Maresca and S. Maslen (ed.), The Banning of Anti-Personnel Landmines: The Legal Contribution of the International Committee of the Red Cross, 1955–1999, Cambridge 2000; L. Doswald-Beck, International humanitarian law and new technology, 106 American Society of International Law: Proceedings of the Annual Meeting 107 (2012).

6 M. Ellen O'Connell, The Power and Purpose of International Law, Oxford 2008; M. Barnett and M. Finnemore, Rules for the World: International Organizations in Global Politics, Ithaca 2004.

7 G. Goertz, P.F. Diehl and A. Balas, The Puzzle of Peace: The Evolution of Peace in the International System, Oxford 2016.

8 D.D. Avant, M. Finnemore and S. K. Sell (ed.), Who Governs the Globe?, Cambridge 2010; J. G. Ruggie, Global governance and "new governance theory": Lessons from business and human rights, 20 Global Governance 5 (2014).

9 C.M. Bailliet and K. Mujezinovic Larsen (ed.), Promoting Peace Through International Law, Oxford 2015.

The First Pillar and the Problems Posed by Lethal Autonomous Weapons

The first pillar, upholding peace and security through the prohibition of the use of force in the conduct of international relations, is anchored in two overarching norms that guide states: peaceful settlement of disputes (through international and regional courts), and international organisations.

The revolution catalysed by lethal autonomous weapons will give rise to two important questions posed to the international community.[10] First, should a lethal autonomous arms race be preventively stopped before it is initiated? Second, should humanity go ahead and allow autonomous weapons to be able to kill without human oversight? The answers to these questions pose a dilemma to international law: how will this revolution in warfare impact the framework that has developed regarding the peaceful settlement of disputes? If the use of unmanned aerial vehicles, known as drones, serves as an indicator of things to come, a few countries are already employing them in situations that could be peacefully settled using a law enforcement framework.[11] The impediments to the use of force set forth in international relations have been eroded to the point where the use of force is employed without legal justification under existing international law.

The use of lethal autonomous weapons during hostilities will only magnify such dilemmas. Therefore, to say that existing international law frameworks will suffice appears complacent to a not-so-distant future which will bring perils that we are already beginning to face, and that we can prevent now. Lethal autonomous weapons will make war easier to pursue. War becomes easier to declare the greater the distance between the attackers and the attacked, as the risk to the attacker is lowered.

The employment of lethal autonomous weapons will mean the lowering of the established thresholds for the use of force, which have been carefully built by international law since the founding of the United Nations. The breakdown of such norms may even lead to an increase in violence between states that do not have autonomous weapons systems. The disintegration of such norms may make fragile conflict situations more precarious. The

10 P. W. Singer, Wired for War, London and New York 2009, p. 14.

11 S. Knuckey, Drones and Targeted Killings: Ethics, Law, Politics, New York 2014; S. Kreps and M. Zenko, The next drone wars, 93:2 Foreign Affairs 68 (2014); S. with J. Kaag, The use of unmanned aerial vehicles in asymmetric conflict: Legal and moral implications, 44:2 Polity 260 (2012).

creation and use of such weapons may also give countries that are technologically superior the feeling that they can use such weapons in disregard of the long established global rules that all actors must abide by.

In sum, the unlawful use of force, or the severe erosion of the existing thresholds for the exercise of force in international relations, will only contribute to weakening the existing system of the rule of law, where solving disputes and problems peacefully should take precedence.

The Second Pillar and the Problems Posed by Lethal Autonomous Weapons

The second pillar, which is efforts to sustain peace and security in the 21st century, is formed by a dense web of global norms that states must operate in. The 22 treaties of a precautionary nature examined here represent the effort to create common global norms. Of particular concern is the question under what conditions new lethal autonomous weapon systems will alter the foundational global norms that comprise the rules of international protection of human rights and international humanitarian law (IHL), and how these weapons will disrupt the regulation of war and conflict under the rules of the UN Charter.

The development of lethal autonomous weapon systems gives rise to concerns from both a human rights and an IHL perspective – the architectures that would be particularly disrupted with the development of such new weapon systems. For both IHL and human rights, the central precepts and requirements are accountability for actions during violent conflict, and the ability to understand the vast number of rules that these branches comprise.[12]

The creation of the United Nations in 1945 accelerated and helped creating the essential framework for peace and security that spawned new global norms for restrictions on war, particularly for the humane treatment of the human being. The existence of new transformative and disruptive lethal weapons systems will undermine these norms and reverse the humanitarian principles codified in the UN Charter.

12 G. Bills, LAWS unto themselves: Controlling the development and use of lethal autonomous weapons systems, 83 George Washington Law Review 176 (2014); D.N. Hammond, Autonomous Weapons and the Problem of State Accountability, 15 Chicago Journal of International Law 652 (2015).

The UN Charter contains some of the foundations of the current global legal order: the prohibition of the use of force, the Security Council as the chief arbiter of peace and security, and the initial contours for the protection of the dignity of the human being as a basis for peace and prosperity in the world. This legal framework gave rise to the 1948 Universal Declaration of Human Rights (UDHR) and the 1949 Geneva Conventions.

The 1966 International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights furnished the legal codification of all human rights initially contained in the UDHR. This human rights framework determines that the right to life is essential, and that its arbitrary deprivation is therefore unlawful under international law. The right to dignity is the basis of all other human rights. Being killed by an autonomous machine, however, is potentially a violation of the right to dignity. Taken together, these two branches of international law, human rights law and international humanitarian law, serve as the key to human moral judgements due to their emphasis on the protection of life.[13]

The implementation of IHL lay dormant until the Vietnam War provided the horrific shock necessary for political action. A conference convened in Tehran in 1968, under the auspices of the UN, provided the impetus for further regulation of internal conflicts with a resolution titled 'Human rights in armed conflicts'. It requested that the General Assembly examine concrete ways to apply the Geneva Conventions to all armed conflicts. As a result, the General Assembly adopted Resolution 2444(XXIII) on 'Respect for human rights in armed conflicts'. This combination of IHL and human rights concerns paved the way for the diplomatic conferences that led to the 1977 Protocols.[14] This contributed further to regulations protecting civilians in war, using Article 3 of the Geneva Conventions as its starting point.[15] This evolution has made IHL a central concern for the international community ever since.[16]

The Security Council played an essential role in the custodianship of IHL after the Cold War came to an end, in its relationship with human rights law. It did so in two ways. One was through the adoption of legally binding resolutions stating that large-scale human rights violations and the resulting humanitarian tragedies are threats to peace and security. The first such resolution was SC/Res 688, passed in 1991 in regards to the Kurds in Northern Iraq. It explicitly referred to the destabilisation caused by the flow of refugees as a threat to inter-

13 C. Heyns, Report of the special rapporteur on extrajudicial, summary or arbitrary executions, UN Doc. A/HRC/23/47, 9 April 2013.

14 D. Schindler, International Humanitarian Law: Its Remarkable Development and its Persistent Violation, 5 Journal of the History of International Law 165 (2003).

15 Ibid.

16 A. Hauptman, Autonomous Weapons and the Law of Armed Conflict, 218 Military Law Review 170 (2013).

national peace and security. The second resolution was SC/Res 770 from 13 August 1992 regarding Bosnia-Herzegovina. Based on Chapter VII of the UN Charter, it called upon all states to facilitate the delivery of humanitarian aid. Several similar resolutions followed in regards to Somalia, Kosovo, Rwanda, Haiti, and Albania. The ones on East Timor and Sierra Leone in 1999 authorised the use of force for humanitarian purposes.[17] The second way the UN Security Council played an essential role was by establishing two ad hoc tribunals: the International Criminal Tribunal for the Former Yugoslavia (ICTY)[18] and the International Criminal Tribunal for Rwanda (ICTR).[19] The UN has also been involved with the Special Court for Sierra Leone, the Extraordinary Chambers in the Courts of Cambodia, and others.

Taken together, these notable changes in the role and action of the Security Council in enacting new international law is a manifest indication of progress in international relations. This new behaviour inaugurates a new era for humankind, with a distinctly humanitarian tenor.[20] The universal global norms protecting human rights through humanitarian law form part of the integral consciousness of humanity and form a common legal code that is broadly adhered to.[21] Seventy years into this new normative order, however, there remain urgent challenges and complexities that need to be addressed. How to best protect the dignity of the individual from the worst atrocities and disrespect of the global norms of humanity remains a primary concern.

It is critical to determine how existing branches of international law apply to protect civilians from harm, and whether what currently exists is sufficient to address the challenges posed by the development of new weapons and technologies. Some authors affirm that existing frameworks are sufficient in principle. In addition, the fact that a particular means or method of warfare is not specifically regulated does not mean that it can be used without restriction. This does not, however, mean that laws cannot be strengthened or that new laws do not need to be created to address future arms and technologies.[22] The question of how to create governance frameworks and what they should include is urgent.

17 SC/Res 1264 (1999); SC/Res 1270 (1999).

18 SC/Res 827 (1993).

19 SC/Res 955 (1994).

20 S. Cardenas, The Geneva Conventions and the Normative Tenor in International Relations, in: S. Perrigo and J. Whitman (ed.), The Geneva Conventions under Assault, London 2010, chapter 1; A.A. Cançado Trindade, International Law for Humankind: Towards a New *Jus Gentium*, Leiden 2013.

21 R. Teitel, Humanity's Law, Oxford 2013; A. Haque, L. Miranda and A. Spain, New Voices I: Humanizing Conflict, 106 American Society of International Law. Proceedings of the Annual Meeting 73 (2012)

22 W. Wallash, A Dangerous Master – How to Keep Technologies from Slipping Beyond Our Control, Philadelphia 2015; E. Prokosch, The Technology of Killing: A Military and Political History of Antipersonnel Weapons, London 1995.

The combined development of human rights law and IHL spawned an era of 'Humanitarian Security Regimes', with universal humanitarian objectives more likely to be effective in galvanising a broader base of support. Concomitantly, international law has evolved to protect individuals and not only states, and a trend towards the 'humanization of law' has emerged.

'Humanitarian security regimes' aim to protect civilians or control or restrict certain weapon systems, and are driven by altruistic imperatives aiming to prohibit harmful behaviour, impede lethal technology, or ban categories of weapons through disarmament treaties. These regimes embrace humanitarian perspectives that seek to prevent civilian casualties, and protect and guarantee the rights of victims and survivors of armed violence. Thus, the chief goals of humanitarian security regimes are to reduce human suffering, to prohibit harm, and to protect victims.

The relevance of the concept of humanitarian security regimes to the weaponisation of artificial intelligence rests on three factors.[23] First, security areas that were previously considered the exclusive domain of states have now been successfully impacted by non-state actors aiming to protect potential and actual victims and survivors, and to reduce human suffering. Demonstrating the growing power of non-state actors, the 1997 Nobel Peace Prize was awarded to Jody Williams and the International Campaign to Ban Landmines (ICBL) in recognition of the new role played by civil society in bringing the treaty to fruition.[24]

Second, states have embraced changes in domains close to their national security (such as armaments), largely in acknowledgement of humanitarian concerns. The Martens Clause was first introduced into international law in 1899 as part of the first Hague Convention, and sets a longstanding and influential normative imperative whereupon states' actions should be driven by ethical and moral concerns.[25] Precisely, it should be used as a guiding map while no regulations are in place.[26] Arguably, the promise of that 1899 provision has

23 D. Garcia, Humanitarian security regimes, 91 International Affairs 55 (2015).

24 J. Borrie, Unacceptable Harm: A History of How the Treaty to Ban Cluster Munitions Was Won, New York and Geneva 2009; J. Borrie, Humanitarian Reframing of Nuclear Weapons and the Logic of a Ban, 90 International Affairs 625 (2014).

25 The importance of the Martens Clause was reaffirmed in ICJ, The Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 8 July 1996.

26 P.W. Singer, The Ethics of Killer Applications: Why Is It So Hard to Talk About Morality When It Comes to New Military Technology?, 9 Journal of Military Ethics 299 (2010).

never been fully reached until now, with the rise of these novel humanitarian security regimes. Third, states are compelled to re-evaluate their national interests and to be duty-bound by a clear humanitarian impetus or reputational concerns.[27]

The key humanitarian principles, now customary, that have been driving disarmament diplomacy in the last century are the prohibition of unnecessary suffering by combatants, the outlawing of indiscriminate weapons, and the need to distinguish between civilians and combatants.[28] Humanitarian concerns have always been part of the equation in multilateral disarmament diplomacy. However, only in recent years have they assumed centre stage and become the driving force, as seen in regard to the 1997 Ottawa Convention on Landmines and the 2008 Convention on Cluster Munitions. Such concerns can also be said to have been at the core of the overwhelming international support for the Arms Trade Treaty.[29]

Finally, it is should be noted that global rule-making, which once was fundamentally anchored on consensus, may indeed be in decay.[30] Consensus seems increasingly outdated, as it no longer reflects the reality and urgency of global governance.[31] For instance, the Arms Trade Treaty abandoned consensus negotiations out of frustration and an inability to create a legally binding document covering legal arms transfers. The Arms Trade Treaty represents a significant shift in negotiation tactics, as it broke free from the constraints of consensus negotiations and instead was shifted to a vote in the UN General Assembly, where approving such treaties requires only a majority.

Sustaining global peace and security norms requires transparency, confidence building, and security mechanisms such as alliances, arms control agreements, nuclear weapons free zones (NWFZs), joint operations, disarmament, conflict resolution, peace-keeping, or reconciliation. Much of this stabilising framework has been set and constructed at the regional level, such as existing NWFZs. The creation of a new weapon system, in the realm of artificial intelligence, will require a totally new expensive political and legal structure. If it is to be like other mechanisms, transparency will be most successful at the regional level.[32]

27 A. Gillies, Reputational Concerns and the Emergence of Oil Sector Transparency as an International Norm, 54 International Studies Quarterly, 103 (2010).

28 J.M. Henckaerts and L. Doswald-Beck (ed), Customary International Humanitarian Law, Cambridge 2005.

29 J.L. Erickson, Stopping the legal flow of weapons: Compliance with arms embargoes, 1981–2004, 50 Journal of Peace Research 159 (2013).

30 N. Krisch, The decay of consent: International law in an age of global public goods, 108 American Journal of International Law 1 (2014).

31 I thank Prof. Robin Geiß for this insight.

32 E. Adler and P. Greve, When security community meets balance of power: overlapping regional mechanisms of security governance, 35 Review of International Studies 59 (2009).

On the global level, transparency could become more problematic with the advent of lethal autonomous weapons.[33] There is a legal responsibility regarding the creation of new weapons under Article 36 of the 1977 Additional Protocol to the Geneva Conventions, which states: "In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party." Therefore, Article 36 mandates states to conduct reviews of new weapons to assess whether they are compliant with IHL. However, only a handful of states carry out weapons reviews regularly, which makes this transparency mechanism insufficient as a tool for creating security frameworks for future arms and technologies.

The Third Pillar and the Problems Posed by Lethal Autonomous Weapons:

The third pillar of peace and security today is the efforts to foster cooperation in cultural, economic, social, and environmental matters that affect all humanity and tackle problems that can only be solved collectively. This framework is based on Article 1(3) of the UN Charter: "To achieve international co-operation in solving international problems of an economic, social, cultural, or humanitarian character, and in promoting and encouraging respect for human rights and for fundamental freedoms for all without distinction as to race, sex, language, or religion", and in the several UN agencies and entities devoted to such goals.

Two recent major achievements by all member states of the United Nations must be highlighted in order to understand how the world can tackle economic, social, and environmental problems in a holistic way: the new United Nations Sustainable Development Goals, and the Paris Agreements on Climate Change. Taken together, they represent a robust map to craft a path for solving some of the worst economic, social and environmental problems facing humanity today.

UN Charter Article 26, which represents a foundational guide, calls on all states, under the auspices and responsibility of the Security Council, to create a system for the regulation of armaments that will focus on the promotion and maintenance of international peace and security. It additionally calls on states to ensure that as little of the world's human and economic resources as possible are used for the creation and maintenance of armaments.

33 H. Roff, The Strategic Robot Problem, 13 Journal of Military Ethics 211 (2014).

United Nations member states have achieved great strides in reaching Article 26's mandate, but much more needs to be done. There is no doubt that establishing such systems regulating arms, which are of course at the heart of states' national security, is costly and demanding. The elaborate legal-political framework to contain the risks, and to avoid the proliferation, of nuclear weapons serves as a useful example.[34] Complying, and ensuring others to comply, has taken an enormous amount of time, and has diverted states' energies and resources. In the case of autonomous lethal weapon systems, the best strategy at this critical juncture for humanity is to ban them, as they would threaten to divert the world's human and economic resources. The UN Charter created a rule of law that has the potential to be an equalizer in international relations.[35] A destabilising weapon system that is deemed to be transformative to the nature of war and peace could create a dual-power structure of those states that are in possession of autonomous weapons on the one hand and those that are not on the other.

Avenues for the way forward

Autonomous lethal weapons present a potentially dangerous challenge to the world and to international law, but there are steps that can be taken. Preventive security governance is the best option for a peaceful future, and based on the principles of international law. Preventive security governance is a strategy to curtail uncertainty regarding the ability to preserve stability and international order.[36] It is the codification of specific or new global norms arising from existing international law that will clarify expectations and universally agreed behaviour concerning a given area of issues. Such an area is characterised either by no rules at all, or by the imprecision of extant rules. The precautionary principle of international law, on whose rationale preventive security governance rests, comprises three components: action to avoid harm regardless of improbability, shifting the burden of proof to supporters of a potentially damaging activity thorough consideration of all alternatives, and transparent decision-making that includes those who would be affected. The precautionary principle calls upon the advocates of actions that may lead to irrevocable damage to take preventive measures to avert harm, even if there is no scientific certainty regarding the danger.

In the coming fifteen years, states should focus all their attention on continuing to construct the architecture of peace. A technology as costly and precariously unpredictable as lethal artificial intelligence, however, would be a distraction from the goals of upholding international security.

34 ICJ, Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 8 July 1996, para. 25.
35 I. Hurd, The International Rule of Law and the Domestic Analogy, 4 Global Constitutionalism 365 (2015).
36 E. Krahmann, Conceptualizing security governance, 38 Cooperation and Conflict 5 (2003).

Historically, several states have embraced efforts to prohibit superfluous and unnecessary armaments. Individual states can become champions of such causes and unleash real progress in disarmament diplomacy. Such champion states have recently helped create extraordinary new international treaties: two prohibiting two classes of superfluous arms – landmines and cluster munitions – and another that created the first treaty regarding the transfer of conventional arms. The 1997 treaty prohibiting mines and the 2008 one that banned cluster munitions were success stories because they prohibited weapons that harmed civilians indiscriminately. The 2013 Arms Trade Treaty, the first global legal agreement on the transfer of conventional arms, is a significant first step to establishing more transparency and accountability in the trade of arms. The presence of an 'epistemic community' – a group of scientists and activists with common scientific and professional language and views that are able to generate credible information – is a powerful tool for mobilizing attention towards action. In the case of autonomous weapons, the International Committee for Robot Arms Control (ICRAC) serves such a purpose. The launch of a transnational campaign is another key element to summon awareness at several levels of diplomatic and global action.[37] The 'Campaign to Stop Killer Robots'[38] is in place and is attracting an unprecedented positive response from around the world.

Conclusions

Another arms race will make everyone less secure and leave the world worse off. Just as nuclear weapons effectively created a dual structure of power in international relations, in the case that lethal autonomous artificial intelligence arises, this new disruptive weapon system will create another parallel system of double-power imbalances.

A preventive ban on lethal autonomous weapons is the vital first step in setting a principled limit on the development of unnecessary and destabilising weapons that have the potential to violate international law.[39] A ban would also be a critical tool in the effort to regulate other robotics and nascent technology that could undermine international peace and security.[40]

37 C. Carpenter, "Lost" Causes: Agenda Vetting in Global Issue Networks and the Shaping of Human Security, Ithaca 2014.

38 See http://www.stopkillerrobots.org/.

39 D. Garcia, The Case against Killer Robots – Why the United States Should Ban Them, Foreign Affairs online, 10 May 2014, www.foreignaffairs.com/articles/141407/denise-garcia/the-case-against-killer-robots.

40 R. Johnson, The United Nations and Disarmament Treaties, UN Chronicle 51, no. 3 (2014): pp. 25-28; J. Altmann, Military Nanotechnology: Potential Applications and Preventive Arms Control, New York 2006, chapters 1-3.

Shared collective action is needed to promote a more secure world, and these efforts should be founded on the gains that have previously been made for national security through the use of pre-emptive action banning dangerous weapons. Prevention instead of reaction is a moral imperative.[41] In the case of lethal autonomous weapons, even if they comply with IHL, but they will have a disintegrating effect on the commonly agreed rules of international law.[42]

Lethal autonomous weapons will make warfare unnecessarily more indiscriminate.[43] They are, additionally, an expensive option, particularly at a time when nations are often looking to cut spending. Further, their creation will create tremendous uncertainty in the world system, and greatly complicate the creation of national defence policies. Countries that develop and deploy such weapons risk losing the ability to shape global discourse in an era that is revolving around increasing regional and international frameworks focused on improving peace and security. Nations today have one of the greatest opportunities in history to promote a better future by creating a protocol that will preventively prohibit the weaponisation of artificial intelligence.

41 P. Lin, Robot ethics: The ethical and social implications of robotics. Cambridge, Mass 2012; P. Lin, Ethical Blowback from Emerging Technologies, 9 Journal of Military Ethics 313 (2010).

42 M.E. O'Connell, 21st Century Arms Controls Challenges: Drones, Cyber Weapons, Killer Robots, and WMDs, 13 Washington University Global Studies Law Review 515 (2014).

43 J. Dill, Legitimate Targets? Social Construction, International Law and US Bombing, Cambridge 2014.

Autonomous Weapons Systems: Risk Management and State Responsibility

Robin Geiß*

Introduction

The development of autonomous weapons systems (AWS) is set to revolutionise weapons technology and military affairs in general.[1] These systems raise a number of challenging ethical, legal and political questions.[2] In the view of this author, as far as 'critical functions' are concerned, meaningful human control must be retained.[3] One particular legal issue that arises with regard to autonomous weapons system is the question of who will be held accountable in the case that something goes wrong and how the risks inherent in this novel and still incompletely understood technology can adequately be managed. These questions are the focus of this chapter.

Accountability challenges in relation to autonomous weapons systems arise because traditional accountability models are typically premised on some form of control and foreseeability. Higher levels of autonomy in weapons systems, however, imply lowered levels of control and foreseeability. Accordingly, the more autonomous a (weapons) system is, the more difficult will it be to establish accountability on the basis of traditional accountability models. This challenge exists with regard to civil uses of autonomous technology (e.g. self-driving cars) in the same way that it exists for military uses of autonomous systems. It follows that a correlation exists between the question of accountability and the notion of 'meaningful human control' (provided states agree to rely on this notion). If the notion of 'meaningful human control' is ultimately defined narrowly – e.g. to require real-time monitoring and human override options – traditional accountability models that are premised on control over specific acts could still function. Conversely, the more broadly the notion of 'meaningful human control' is defined, the stronger the case for a different model.

* Professor of International Law and Security, University of Glasgow, UK. This contribution builds on the statement delivered to the Third CCW meeting of experts on lethal autonomous weapons systems, Geneva, 11-15 April 2016, available at: http://www.unog.ch/80256EDD006B8954/%28httpAssets%29/00C95F16D6FC38E4C1257F9D0039B84D/$file/Geiss-CCW-Website.pdf.

1 P. Singer, Wired for War, New York 2009, p. 179 et seq.

2 R. Geiß, The International Law Dimension of Autonomous Weapons Systems, Friedrich-Ebert-Stiftung, Study, October 2015, http://library.fes.de/pdf-files/id/ipa/11673.pdf.

3 Ibid.

This, however, does not mean that there is an inevitable or insurmountable 'accountability gap'.[4] Especially in the area of state responsibility – the conceptual challenges are greater when focusing on individual criminal responsibility – accountability challenges can be overcome by way of regulation and clarification of existing laws. There is no conceptual barrier for holding a state (or individual human being) accountable for wrongful acts committed by a robot or for failures regarding risk minimisation and harm prevention. There is therefore no need to devise a new legal category of 'e-persons', 'non-human actors', or 'virtual legal entities', and any idea that a robot itself could or should be held accountable should be rejected.[5]

Article 36-weapons review as an important first step towards risk mitigation

It is clear and undisputed that autonomous weapons systems that cannot comply with the laws of armed conflict must not be fielded. Article 36 Additional Protocol I (API) is therefore a logical starting point for any discussions about autonomous weapons systems.[6] It is important to note, however, that even weapons that have been thoroughly reviewed may fail or malfunction in combat. This is true for any type of weapon. Accidents can happen. There is no such thing as absolute certainty or failure-proof technology. With regard to autonomous weapons systems, however, there is concern that because of their autonomy and the complex software they involve – even after a thorough weapons review in the sense of Article 36 API – there remains a higher than usual risk and degree of unpredictability as to how exactly they will operate under actual battlefield conditions and a resultant higher risk of accidents and wrongful conduct.

There is considerable controversy among states as to how this issue of residual risks and unpredictable robotic activity should be approached. Some have argued that because of these risks, autonomous weapons systems should be banned altogether. Others seem to hold the view that residual risks, i.e. risks that remain after the required Article 36 API weapons review, are generally acceptable. In the view of this author, both approaches go

4 C. Heyns, Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, UN Doc. A/HRC/23/47 (2013), para. 80: "If the nature of a weapon renders responsibility for its consequences impossible, its use should be considered unethical and unlawful as an abhorrent weapon."

5 See also M. Sassòli, Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified, 90 International Law Studies, 308 (2014) 322.

6 See only e.g. Statement on the Implementation of Weapons Reviews under Article 36 Additional Protocol I by Germany, 13-17 April 2015, available at: http://www.unog.ch/80256EDD006B8954/%28httpAssets%29/CBB705417D436311C1257E290046EF14/$file/2015_LAWS_MX_Germany_IHL.pdf.

too far. Article 36 API should be understood as an important first step in the process of risk mitigation. But further fine-tuning and additional risk mitigation measures are required. Article 36 API functions as a 'door-opener' for a newly developed weapons technology that passes the relevant threshold of scrutiny to be deployed on the battlefield. However, the threshold imposed by Article 36 API is relatively low. Further clarification of the requirements of an Article 36-weapons review notwithstanding, Article 36 API only prohibits weapons systems that cannot abide by international law in *any* realistic deployment scenario. In other words, it (only) rules out the 'worst' types of weaponry. This means, however, that even high-risk military technology may be permitted on the battlefield and there may thus be – and in the case of autonomous weapons system and in light of the limited testability of the complex software they involve there clearly is[7] – an urgent need for additional harm prevention and risk mitigation measures.

Towards a more comprehensive risk mitigation and harm prevention regime for autonomous weapons systems

In the case of autonomous weapons systems such a risk mitigation model should be based on the following rationale: The deployment of autonomous weapons systems is not per se unlawful but it is a high-risk activity. This novel technology – especially if it is used in a complex, dynamic battlefield environment – is not (yet) fully understood. There is at best predictable unpredictability. It follows that a state that benefits from the advantages of this high-risk technology, i.e. the various (strategic) gains associated with autonomous weapons systems,[8] should be required to mitigate and control these risks as far as possible and should be held responsible whenever the (unpredictable) risks inherent in this technology are realised. Instead of focusing (only) on state responsibility relating to wrongful conduct on the battlefield, the focus should thus be shifted to also include risk mitigation and harm reduction obligations at the development pre-deployment stages and state responsibility arising from failure to abide by these obligations.

7 See presentation by M. Hagstrom delivered to the Third CCW meeting of experts on lethal autonomous weapons systems, Geneva, 11-15 April 2016, available at: http://www.unog.ch/80256EDD006B8954/%28httpAssets%29/8121CB0B41813E6CC1257F950027AAE0/$file/2016_LAWS+MX+Presentations_ChallengestoIHL_Martin+Hagstr%C3%B6m.pdf.

8 See e.g. M.N. Schmitt and J.S. Thurnher, Out of the Loop: Autonomous Weapons Systems and the Law of Armed Conflict, 4 Harvard National Security Journal 231 (2013).

On the basis of this rationale, a state could be held responsible principally in two different ways, namely for failures regarding risk prevention and harm reduction including at the development and pre-deployment stage (see below 1), as well as for specific wrongful actions of the autonomous weapons system (see below 2). In other words, a combination of due diligence obligations aimed at harm prevention and risk mitigation on the one hand, and state responsibility (potentially to be extended through the imposition of stricter burdens of liability, provided states were willing to agree on such a model) for failures to abide by specific rules governing the conduct of hostilities on the other, could help to absorb some of the risks inherent in robotic activity on the battlefield.

1 Due diligence obligations regarding risk mitigation and harm prevention

Prevention is always better than cure. Due diligence obligations aiming at risk prevention and harm reduction at the development, pre-deployment and deployment stage will limit the risk of wrongful conduct from the outset. With respect to the Laws of Armed Conflict, such obligations could *inter alia* be derived from common Article 1 GC I-IV (and corresponding customary international law), which requires states to *ensure respect* for the laws of armed conflict in all circumstances.[9] The problem is therefore not the lack of a legal basis but the lack of clarity as to what exactly it is that the due diligence obligation to *ensure respect* would require with regard to autonomous weapons systems. As is well known, due diligence obligations require what a reasonable actor would do under similar circumstances.[10] The problem with autonomous weapons systems is that it is hard to know what is considered reasonable when dealing with a new technology for which clear standards, practical experiences and established benchmarks do not (yet) exist. Of course, due diligence obligations and the standard of reasonableness will always leave states a considerable margin of discretion regarding the implementation of the obligation. But without any clarification and in light of the many uncertainties still surrounding this novel technology, due diligence obligations aimed at mitigating the risks posed by autonomous weapons systems could remain empty shells.

9 R. Geiß, The Obligation to Respect and Ensure Respect for the Conventions, in: A. Clapham, P. Gaeta and M. Sassoli (ed.), The 1949 Geneva Conventions – A Commentary, Oxford 2015.

10 D. French and T. Stephens, ILA Study Group on Due Diligence in International Law, 1st Report, http://www.ila-hq.org/en/committees/study_groups.cfm/cid/1045.

It is therefore recommended that in addition to the clarification of Article 36 API, emphasis should also be put on the specification and clarification of due diligence obligations aimed at risk prevention and harm reduction. There are various ways in which risks resulting from unpredictable robotic activities could be mitigated. Alternatively or cumulatively, this could be achieved for example by way of implementing domestic laws requiring strict controls at the development and pre-deployment stage, in-built automatic deactivation devices, real-time monitoring, or conservative programming, i.e. a requirement to 'shoot second', to stand-down or double-check in case of doubt. Indeed, given that robots engaged in hostilities – unlike human soldiers – are not at risk of losing their lives, programming them in a way that is more restrictive than what IHL normally permits, could be an option to mitigate the risk of overly violent acts from the outset.

Of course, due diligence obligations are relative in the sense that they may require different activity depending on capacity, the specific circumstances of a given case, and the level of risk involved.[11] The Commentary to Article 3 of the Draft Articles on the Prevention of Transboundary Harm confirms that the due diligence standard should be "appropriate and proportional to the degree of risk".[12] Thus, as a general rule, the higher the risk, the stricter the obligation to mitigate risks. There is thus a graduated set of risk mitigation obligations depending on deployment scenarios, the range of tasks to be fulfilled, and the specific features of the weapons system at issue. It follows that risk mitigation obligations will be rather low when a robot is deployed in a low-risk scenario, e.g. a predetermined area of operation (e.g. a submarine or outer-space context) where no human beings are present and where the robot is tasked to neutralise a specific military object (e.g. a naval mine). Conversely, if a robot were to be deployed in a complex, highly dynamic urban environment with a resultant significant risk of detrimental impact on human beings and civilian infrastructure, risk mitigation obligations would be very high. It is noteworthy that in this context the UN Guiding Principles on Business and Human Rights confirm that due diligence requirements increase in situations in which the risks of harm are known to be particularly significant.[13]

11 Seabed Mining Advisory Opinion, 50 ILM 458 (2011), para. 117.

12 International Law Commission, Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, UN GAOR 56th Sess., Supp. No. 10, UN Doc. A/56/10 (2001), Commentary to Article 3, para. 11.

13 Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework, UN Human Rights Council, UN Doc. A/HRC/17/31 (21 March 2011). In particular, principle 17(b) explicitly states that due diligence "will vary in complexity with (...) the risk of severe human rights impacts".

2 State responsibility

Risk mitigation and harm prevention is one question; another question is what happens if something goes wrong on the battlefield. In accordance with general rules on state responsibility[14], a state is responsible for all violations of international (humanitarian) law that are attributable to it. In other words, state responsibility is premised on two central elements: attribution of conduct and a violation of international law.

a Attribution of conduct

No particular legal challenges arise with regard to the attribution of acts committed by autonomous weapons systems. For as long as human beings decide on the deployment of these systems, accountability can be determined on the basis of the established rules on attribution.[15] It is a long-standing rule of customary international law, set forth in Article 3 of the 1907 Hague Convention (IV) and repeated in Article 91 API that states are responsible for their state organs, and that includes responsibility for "all acts committed by persons forming part of its armed forces". This rule also applies by virtue of customary international law as confirmed in the ICRC Customary Law Study according to which: "[a] State is responsible for violations of international humanitarian law attributable to it, including: (a) violations committed by its organs, including its armed forces."[16] Thus, if a member of the armed forces (i.e. a state organ) of state A decides to deploy a robot on a combat mission, all activities carried out by the robot are attributable to that state. The mere fact that a weapons system has (some) autonomous capabilities does not alter this assessment.

b The commission of an internationally wrongful act

The determination whether an internationally wrongful act has been committed, i.e. whether a (primary) norm of international law has been violated by an autonomous weapons system, can in some cases be more problematic. Many rules of international (humanitarian) law are violated whenever their objective requirements are met. As the ILC

14 Articles on Responsibility of States for Internationally Wrongful Acts, Report of the International Law Commission, 53rd Session, UN Doc. A/56/10 (2001), 43–59 (noted in GA Res. 56/83, 12 December 2001, UN Doc. A/RES/56/83 (2001)). According to Article 1: "Every internationally wrongful act of a State entails the international responsibility of that State."

15 Ibid., Art. 4-11.

16 ICRC Customary Law Study, Rule 149.

Commentary explains: "[i]n the absence of any specific requirement of a mental element in terms of the primary obligation, it is only the act of a State that matters, independently of any intention."[17] In this case no particular challenges arise. Whenever such a rule is (objectively) violated by a robot (and provided the activity is attributable to a state, see above a)), state responsibility arises. Thus, if one takes the view that most or all of the prohibitions contained in IHL do not contain such an 'intent' or 'fault' requirement, establishing state responsibility for wrongful conduct of autonomous weapons systems will be a relatively straightforward exercise. For example, with regard to APII it was recently held that: "[n]otably, there is no 'intent' requirement in relation to the prohibitions under APII, meaning that State responsibility is to be assessed 'objectively' rather than 'subjectively': the intent or advertence of relevant State organs or agents is not relevant to an assessment of whether a violation of APII has occurred."[18]

Some primary rules of international (humanitarian) law, however, in order to be violated, may require an element of 'fault' (negligence, recklessness, or intent). Which rules belong to the first or the second category depends on the specific requirements and interpretation of the primary humanitarian law rule in question.[19] If the rule in question belongs to the second category, i.e. if it is a rule that requires an element of 'fault' in order to be violated, it may indeed be difficult or impossible to establish state responsibility for robotic activity 'violating' that rule.

In certain scenarios, namely when the military commander acts with the clear intent to violate a rule of international humanitarian law, even these scenarios will not pose any problems. Thus, if in the course of an armed conflict a military commander intentionally programs an autonomous weapons system to attack civilians, it is clear that the commander is individually responsible for having committed a war crime and that the state of

17 ILC Commentaries, Article 2, para. 10.

18 P. Sands et al., The Lawfulness of the Authorization by the United Kingdom of Weapons and Related Items for Export to Saudi Arabia in the Context of Saudi Arabia's Military Intervention in Yemen, 11 December 2015, para. 2.16., available at: http://www.saferworld.org.uk/resources/view-resource/1023-the-lawfulness-of-the-authorisation-by-the-united-kingdom-of-weapons-and-related-items-for-export-to-saudi-arabia-in-the-context-of-saudi-arabias-military-intervention-in-yemen.

19 According to the ILC Commentaries: "Whether there has been an internationally wrongful act depends, first, on the requirements of the obligation which is said to have been breached"; ILC Commentaries, Article 1, para. 1. The ILC Commentaries further state: "Whether responsibility is "objective" or "subjective" in this sense depends on the circumstances, including the content of the primary obligation in question. The articles lay down no general rule in that regard. The same is true of other standards, whether they involve some degree of fault, culpability, negligence or want of due diligence. Such standards vary from one context to another for reasons which essentially relate to the object and purpose of the treaty provision or other rule giving rise to the primary obligation"; Article 2, para. 3.

which the commander is a state organ is likewise responsible for a violation of the laws of armed conflict.[20] In essence, there is no difference between a member of the armed forces shooting a civilian or a soldier programming a robot to shoot at a civilian. But whereas this clear-cut 'intent scenario' does not raise any particular legal challenges, other – presumably more common – scenarios involving autonomous weapons systems could raise considerable accountability challenges (in such cases in which the violated rule in question contains an 'intent requirements').

The following scenario may help to illustrate the problem: A military commander may field a thoroughly tested and duly authorised autonomous weapons system, which – because it operates autonomously in a complex and dynamic battlefield environment – nevertheless unexpectedly violates the laws of armed conflict. There is no indication that the military commander acted with intent or negligence. And given that intent and negligence denote human mental states, they are by definition absent in a robot. What is more, given the complexity of these systems it may in any case be difficult to prove what exactly went wrong. As a consequence, if the primary rule in question is one that requires an element of fault, it may be impossible to establish or prove state responsibility.

c A stricter liability regime to overcome potential accountability gaps?

In order to address this particular challenge, a (future) liability regime for autonomous weapons systems could potentially be designed so as to not require any proof of fault ('strict liability'), or to reverse the burden of proof ('presumed liability'). From the outset, however, it should be pointed out that the prospects for such a liability regime in the domain of the laws of armed conflict are slim. Strict liability (also known as 'absolute liability', 'operator's liability', or *'responsabilité pour risqué crée'*) means that the issue of 'fault' (negligence, recklessness, intent) is removed from the consideration. Responsibility is triggered automatically whenever the risks inherent in unpredictable robotic behaviour are realised. Under such a strict liability regime, it is irrelevant what the programmer, operator, commander in charge, or state behind the operation thought or expected what the robot might to. All that matters is what the autonomous weapons system actually did. Whether this was due to

20 Articles on Responsibility of States for Internationally Wrongful Acts (n 15), Article 7: "The conduct of an organ of a State or of a person or entity empowered to exercise elements of the governmental authority shall be considered an act of the State under international law if the organ, person or entity acts in that capacity, even if it exceeds its authority or contravenes instructions."

technical failure of its sensors, unforeseen (external) interference, changing environmental conditions, programming errors, other (software) defects, or the autonomously calculated outcome of the robot's algorithms, makes no difference.

Strict liability regimes are not uncommon when dealing with hazardous activity and processes of high complexity that may make it inherently difficult to identify and prove what exactly went wrong. On the domestic level, product liability regimes often involve strict liability. In international law, apart from the ILC Draft Principles on the Allocation of Loss in the Case of Transboundary Harm Arising out of Hazardous Activities,[21] the so-called Outer Space Treaty 1967 and the Space Liability Convention 1972 – certain issues of interpretation notwithstanding – are relevant cases in point. According to Article VII of the Outer Space Treaty, "[e]ach State Party to the Treaty that launches [...] an object into outer space [...] is internationally liable for damage to another State Party to the Treaty [...]."[22] The considerations that led to the adoption of this liability regime are in many ways similar to the considerations regarding autonomous weapons systems. At the time of drafting and adoption of the Outer Space Treaty – i.e. during the 1960s – outer space technology was considered to be incompletely understood and perceived as a high-risk and highly complex endeavour with unpredictable outcomes. The very same rationale applies to autonomous weapons systems today. Unsurprisingly, strict liability regimes are currently also being discussed (domestically) with respect to civil uses of autonomous technology. The Swedish automaker Volvo recently pledged to be 'fully liable' for accidents caused by its self-driving technology.[23]

In addition to overcoming specific accountability challenges associated with high-risk, unpredictable and highly complex activity, a strict international liability regime would create strong incentives to deploy autonomous weapons systems cautiously, to take weapons

21 According to Principle 4 ('Prompt and adequate compensation') of the ILC Draft Principles on the Allocation of Loss in the Case of Transboundary Harm Arising out of Hazardous Activities: "Each State should take all necessary measures to ensure that prompt and adequate compensation is available for victims of transboundary damage caused by hazardous activities located within its territory or otherwise under its jurisdiction or control. These measures should include the imposition of liability on the operator or, where appropriate, other person or entity. *Such liability should not require proof of fault*" (emphasis added).

22 This provision was further refined and specified by virtue of the 1972 Liability Convention. But with respect to damage caused by a space object on the surface of the Earth or an aircraft in flight, liability remains absolute (Article II 1972 Liability Convention).

23 See statement delivered by Volvo President and CEO Håkan Samuelsson, available at: http://www.volvocars.com/intl/about/our-innovation-brands/intellisafe/intellisafe-autopilot/news/volvo-cars-responsible-for-the-actions-of-its-self-driving-cars#.

reviews and steps towards risk minimisation and harm prevention seriously, to program autonomous weapons systems 'conservatively', and to implement strict product liability regimes domestically.

Further reflection, however, is required as to how such a liability regime could be applied in the context of an armed conflict and with regard to systems that are by definition designed to lawfully cause certain damage. The liability regimes laid out in the Outer Space Treaty and Space Liability Convention, whereby a state shall be absolutely liable for (any) damage caused by its space objects on the surface of the earth, are obviously not directly transferable to autonomous weapons systems. Autonomous weapons systems would require a more nuanced liability regime specifically tailored to the context of an armed conflict. After all, the causation of certain damages (e.g. the destruction of a military objective) is clearly permissible in times of armed conflict and does not lead to any state responsibility. What is more, the risks and unpredictability associated with autonomous weapons systems might be more relevant in certain scenarios (e.g. when the system is deployed in an urban area) than in others (e.g. when the system is deployed in a naval context or in outer-space). One could therefore envision a graduated liability regime whereby strict or presumed liability is imposed in certain scenarios or with respect to certain, fundamental rules, but not in all scenarios and not for the causation of damage generally, or in relation to the entire body of rules comprising the laws of armed conflict. Such a graduated liability regime that combines strict liability with other forms of liability may be best suited to respond to the different risks and uncertainties inherent in autonomous weapons systems that may be deployed in vastly different contexts.

Conclusion

Notably, such a liability regime does not currently exist. And it is highly doubtful that states would be willing to agree on such a regime with respect to a weapons system and in the domain of the laws of armed conflict. This makes it all the more important that in addition to ongoing discussions about the notion of 'meaningful human control' and the specification of the requirements of an Article 36 API weapons review, due diligence obligations regarding risk mitigation and harm prevention are further elaborated and specified so as to better absorb the risks associated with autonomous weapons systems.

Legal Review of New Weapons, Means and Methods of Warfare

Gilles Giacca *

Introduction

As rapid advances continue to be made in new and emerging technologies of warfare, notably those relying on information technology and robotics, it is important to ensure informed discussions on the many and often complex challenges raised by these new developments. Although new technologies of warfare, such as autonomous weapons systems,[1] are not specifically regulated by international humanitarian law treaties, their development and employment in armed conflict does not occur in a legal vacuum. As with all weapons systems, they must be capable of being used in compliance with international humanitarian law (IHL), and in particular its rules on the conduct of hostilities. The responsibility for ensuring this rests, first and foremost, with each state that is developing these new technologies of warfare.

In accordance with Article 36 of Additional Protocol I (API), each State Party is required to determine whether the employment of a new weapon, means or method of warfare that it studies, develops, acquires or adopts would, in some or all circumstances, be prohibited by international law. Legal reviews of new weapons, including new technologies of warfare, are a critical measure for states to ensure respect for IHL. The law anticipated to a certain extent advances in weapons' technology and the development of new means and methods of waging war, and Article 36 is clear evidence of that anticipation.

* Dr, Legal Adviser, Arms Unit, Legal Division, International Committee of the Red Cross.

1 The ICRC has defined autonomous weapons systems as "[a]ny weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention". ICRC, Autonomous weapon systems technical, military, legal and humanitarian aspects, Report of an Expert Meeting held 26-28 March 2014, November 2014, https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf ; ICRC, International humanitarian law and the challenges of contemporary armed conflicts, Report to the 32nd International Conference of the Red Cross and Red Crescent held 8-10 December 2015, October 2015, https://www.icrc.org/en/download/file/15061/32ic-report-on-ihl-and-challenges-of-armed-conflicts.pdf, pp. 44-47; see also ICRC, Autonomous weapon systems: Implications of increasing autonomy in the critical functions of weapons, Report of an Expert Meeting held 15-16 March 2016, 2016, https://shop.icrc.org/autonomous-weapon-systems.html?___store=default.

The present chapter will address only a selection of some of the key legal and policy questions associated with weapons reviews in light of the current debate on autonomous weapons systems. It will first discuss the importance of conducting legal reviews. It will then examine the legal review's scope and functional aspects, as well as some of the specific challenges posed by autonomous weapons systems.

Why is conducting legal reviews important?

Setting or improving legal reviews procedures is important for a number of reasons.

First, most of the State Parties to the CCW are party to API. This means that they are legally required to comply with the requirements of Article 36 of that Protocol. It is arguable that a duty to conduct weapons reviews is also derived from the general obligation under Common Article 1 to the four Geneva Conventions to ensure respect for IHL. This obligation would require High Contracting Parties to ensure that their new weapon, means or method of warfare can be used in accordance with IHL.[2] It is self-evident that proceeding to such a review contributes to ensuring that a state's armed forces are capable of conducting hostilities in accordance with its international obligations. With regard to customary international humanitarian law, from the currently available evidence it is unclear whether the obligation to conduct legal reviews of weapons, means and methods of warfare is of a customary law nature.[3]

Second, and related to the previous point, reviewing the legality of new weapons also makes sense as a matter of policy. It is in each state's interest, regardless of whether it is party to API, to assess the lawfulness of its new weapons in order to ensure that it is able to comply with its international legal obligations during armed conflicts, and that new weapons are not employed prematurely under conditions in which compliance with IHL cannot be guaranteed. This can be especially important in light of rapid development of new weapons technologies, and it would give the opportunity to the state to develop its own national expertise in law and weapons.

2 I. Daoust et al., New Wars, New Weapons? The Obligation of States to Assess the Legality of Means and Methods of Warfare', 84 International Review of the Red Cross 352 (2002).

3 Some argue that it is customary, see W.H. Boothby, Weapons and the Law of Armed Conflict, Oxford 2009, pp. 341-342.

Third, promoting, wherever possible, exchange of information and transparency in relation to weapons review mechanisms and procedures can enhance confidence and trust among states, confidence-building being among one of the purposes of the CCW.[4] Moreover, in light of Article 84 API, it can be submitted that there is a requirement for states to share among each other their review procedures in order to build confidence that new weapons comply with the existing law.[5]

On the question of how well the commitments under Article 36 are being implemented, it is fair to say that despite this legal requirement and the large number of states that develop or acquire new weapons systems every year, only a small number are known to have formal mechanisms in place to carry out legal reviews of new weapons. Further efforts are therefore needed to implement this obligation by states. One of the reasons for poor implementation may be that some states assume that when they acquire certain weapons they can safely rely either on the manufacturers' testing or on the reviews conducted by states from which they are procuring the weapons. This is disputable, since obligations under international law differ between states, and even when they are subject to the same obligations, there are often differences in interpretation and implementation. Hence, it is important for states to conduct their own weapons review.

Different views have been expressed on the adequacy of legal reviews of new weapons for ensuring IHL compliance of autonomous weapons systems, especially given the apparent low level of implementation among states, and the possibility of inconsistent outcomes of national legal reviews. One can say that this is not different from other rules of international law; the challenge remains national implementation. This is why states gather regularly in multilateral forums like the CCW to share their views and practice on how they interpret and implement their international obligations. It is indeed an opportunity for states to share their experiences on legal reviews in order to create confidence that the unique questions and challenges raised by autonomy in weapons systems are also dealt with on the domestic level. The CCW could be an appropriate forum to share and learn good practice between states and see how other states implement their obligations.

4 Sixth preambular paragraph of the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects as amended on 21 December 2001(CCW).

5 Art. 84 states that "[t]he High Contracting Parties shall communicate to one another, as soon as possible, through the depositary and, as appropriate, through the Protecting Powers, their official translations of this Protocol, as well as the laws and regulations which they may adopt to ensure its application".

In addition, it is submitted that the question is not so much whether national legal reviews are sufficient or not to deal with new technologies of warfare such as autonomous weapons systems. The need to conduct legal reviews is a legal obligation and remains a critical means for states to ensure respect for IHL regarding any type of weapons. Having said that, it is fair to say that legal reviews do not provide all the answers and efforts to strengthen national review procedures, so they should be seen as complementary and mutually reinforcing multilateral discussions on autonomous weapon systems.

On the question of transparency, it is true that there is little published information available about states' weapon procurement and review processes for a number of reasons: commercial, military, and national security concerns, since such reviews often deal with classified material relating to the performance and use of weapons. However, states are encouraged to share information, to the extent possible, on their legal review mechanisms, i.e. on their procedures to review new weapons. On the one hand, disclosing the process can be seen as a way to show state commitment to legal compliance, and to set the example of what responsible militaries do. On the other, this could foster the development of common standards or best practices for weapons review in the longer run.

What is the scope of the review and how should it be done?

Over the years, the ICRC has consistently promoted the importance of conducting legal reviews of new weapons. In order to provide a tool to assist states in establishing weapons review mechanisms, in 2006 the ICRC drew up 'A Guide to the Legal Review of New Weapons, Means and Methods of Warfare', which it prepared in consultation with close to thirty military and international law experts, including government experts.[6] The Guide aims to assist states in establishing or improving national procedures to determine the legality of new weapons being developed or acquired. The Guide also provides the ICRC's interpretation of what is required and its recommendations for what a review mechanism should look like, based notably on the ICRC Commentary to Article 36 and existing practice of states.[7]

6 ICRC, A Guide to the Legal Review of New Weapons, Means and Methods of Warfare Measures to Implement Article 36 of Additional Protocol I of 1977, 2006, available at: https://www.icrc.org/eng/resources/documents/publication/p0902.htm.

7 At the time of writing, the ICRC is gathering new information from actual practice in order to update this Guide.

This Guide highlights both the issues of substance and those of procedure to be considered in establishing a legal review mechanism. However, Article 36 does not prescribe a method or format for weapons reviews. States are left with a wide margin of appreciation in the domestic implementation of this rule.

The legal review should apply to weapons in the widest sense as well as the ways in which they are used, bearing in mind that a means of warfare cannot be assessed in isolation from its expected method of use.[8] Thus, the legality of a weapon does not depend solely on its design or intended purpose, but also on the manner in which it is expected to be used on the battlefield. According to the ICRC Guide, a weapon used in one manner may pass the Article 36 test, but may fail it when used in another manner. This is why Article 36 requires a state "to determine whether its employment would, *in some or all circumstances*, be prohibited" by international law.[9]

The existing law that determines the legality of new weapons technology includes:

- specific rules of international law prohibiting or restricting the use of specific weapons (e.g. the Biological Weapons Convention (BWC), the Convention on Certain Conventional Weapons (CCW), landmines, cluster munitions)
- general rules of IHL applicable to the use of weapons, including:
  - whether the weapon is of a nature to cause superfluous injury or unnecessary suffering;
  - whether the weapon is likely to have indiscriminate effect;
  - whether the weapon is expected to cause widespread, long-term and severe damage to the natural environment;
  - whether the weapon is likely to be affected by future developments in the law;
  - prohibitions or restrictions based on the principles of humanity and the dictates of public conscience.[10]

The assessment of a weapon in light of the relevant rules will require an examination of all relevant empirical information pertaining to the weapon, such as its technical description and actual performance, and its effects on health and the environment. This is the rationale

8 ICRC, A Guide to the Legal Review of New Weapons, Means and Methods of Warfare Measures to Implement Article 36 of Additional Protocol I of 1977, 2006, available at: https://www.icrc.org/eng/resources/documents/publication/p0902.htm.

9 ICRC Guide (n 6), p.10.

10 For a detailed list of rules to be applied to new weapons, means and methods of warfare, see ibid., pp. 10-17.

for the involvement of experts of various disciplines in the review process. Multidisciplinary expertise is important in order to be able to understand how the weapon functions, its capabilities, and its limitations, and more generally to understand the technology itself.[11]

New weapons, means and methods of warfare include weapons in the broadest sense and ways in which weapons are used. According to the ICRC Guide, the rule covers:

- weapons of all types – be they anti-personnel or anti-materiel, 'lethal', 'nonlethal', or 'less lethal' – and weapons systems;
- the ways in which these weapons are to be used pursuant to military doctrine, tactics, rules of engagement, operating procedures, and countermeasures;
- all weapons to be acquired, be they procured further to research and development on the basis of military specifications, or purchased 'off the shelf';
- a weapon which the state is intending to acquire for the first time, without necessarily being 'new' in a technical sense;
- an existing weapon that is modified in a way that alters its function, or a weapon that has already passed a legal review but that is subsequently modified;
- novel uses of existing capabilities or equipment;
- an existing weapon where a state has joined a new international treaty which may affect the legality of the weapon.[12]

Concerning the functional aspects of the review mechanisms, the ICRC Guide provides a number of elements on how, for instance, the mechanism should be established or what the structure and composition of the mechanism should be. At the minimum, there should be a formal standing mechanism or procedures ready to carry out reviews. It should be mandatory and take place in a systematic way. It is key that the review process begins at the earliest possible stage of the procurement process (study, development, acquisition, adoption), and that it applies a multidisciplinary approach.[13]

11 A. Backstrom and I. Henderson, New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews, 94 International Review of the Red Cross 483 (2012).

12 ICRC Guide (n 6), pp. 9-10.

13 Ibid., pp. 20-28.

What are the specific challenges posed by autonomous weapons systems?

Weapons reviews face certain practical challenges regarding the assessment of whether an autonomous weapons system will perform as anticipated in the intended or expected circumstances of use.[14] Taking human beings out of the critical functions of selecting and attacking targets raises important questions, including how 'targeting rules' (e.g. the rules of proportionality and precautions in attack) can be considered at the weapons review stage, before the weapon system has been deployed. Thus, where it is the weapon that takes on the targeting functions, the legal review would demand a very high level of confidence that the weapon is capable of carrying out those functions in compliance with IHL. The decision to deploy and use a particular weapon by the commander or operator can be based on constraints or parameters concerning its use, which are developed in the weapons review. Those are generally integrated into the military instructions or guidelines, for instance to limit the use to a specific environment or situation.

Key questions include whether the weapons system would function in a way that respects the obligation to distinguish military objectives from civilian objects, combatants from civilians, and active combatants from persons *hors de combat*. Another question is whether a weapons system would function in a way that respects the obligation to weigh the many contextual factors and variables to determine whether the attack may be expected to cause incidental civilian casualties and damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated, as required by the rule of proportionality. A further question is whether the weapons system could function in a way that respects the obligation to cancel or suspend an attack if it becomes apparent that the target is not a military objective or is subject to special protection, or that the attack may be expected to violate the rule of proportionality, as required by the rules on precautions in attack.

For autonomous weapons systems intended for use in contexts where they are likely to encounter protected persons or objects, there are serious doubts as to whether they would function in a way that respects the obligation to carry out the complex, context-dependent assessments required by the IHL rules of distinction, proportionality and precautions in attack.[15] These are inherently qualitative assessments in which unique human reasoning and judgement will continue to be required.

14 See generally ICRC 2016 (n 1).

15 See generally Henckaerts & Doswald-Beck, Customary International Humanitarian Law (ICRC/CUP: 2005), Rules 1-24, available at: https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul.

The above challenges for IHL compliance will need to be carefully considered by states when carrying out legal reviews of any autonomous weapons system they develop or acquire. As with all weapons, the lawfulness of a weapon with autonomy in its critical functions depends on its specific characteristics, and whether, given those characteristics, it can be employed in conformity with the rules of IHL in all of the circumstances in which it is intended and expected to be used. The ability to carry out such a review entails fully understanding the weapon's capabilities and foreseeing its effects, notably through testing. Yet foreseeing such effects may become increasingly difficult if autonomous weapons systems were to become more complex or be given more freedom of action in their operations, and therefore become less predictable.

Predictability about the operation of an autonomous weapons system in the context in which it is to be used must be sufficiently high to allow an accurate legal review. Indeed, deploying a weapons system whose effects are wholly or partially unpredictable would create a significant risk that IHL will not be respected. The risks may be too high to allow the use of the weapon, or else mitigating the risks may require limiting or even obviating the weapons' autonomy.

An additional challenge for reviewing the legality of an autonomous weapons system is the absence of standard methods and protocols for testing and evaluation to assess the performance of these weapons, and the possible risks associated with their use. Questions arise regarding: how is the reliability (e.g. risk of malfunction or vulnerability to cyber-attack) and predictability of the weapon tested? What level of reliability and predictability are considered to be necessary? The legal review procedure faces these and other practical challenges to assess whether an autonomous weapons system will perform as anticipated in the intended or expected circumstances of use. It is hoped that states and military experts will address some of these key questions in the future debates on autonomous weapons systems.

Conclusion

The recognition by states of the importance of reviewing new weapons to ensure their compatibility with international law in the CCW discussions on autonomous weapons systems is a positive outcome. This should be seen as an aspect that is complementary to other debated questions, such as human control or consideration of human-machine interaction, which may provide a useful baseline from which common understandings among states can be developed.

The forthcoming Fifth Review Conference of the CCW (12 to 16 December 2016) is an important moment for States Parties to examine the status and operation of the Convention and its Protocols, to assess developments that have occurred in the use of weapons and weapons technology, and to consider enhancing the protections of international humanitarian law for the benefit of civilians and combatants. Already in 2006, the Third Review Conference of the CCW urged states that do not already do so to conduct legal reviews of new weapons, means or methods of warfare.[16] The Fifth Review Conference, as well as future discussions on autonomous weapons systems, presents another opportunity for states to consider the importance of carrying out timely legal reviews of newly developed or acquired weapons.

16 In paragraph 17, states solemnly declared "[t]heir determination to urge States which do not already do so to conduct reviews to determine whether any new weapon, means or methods of warfare would be prohibited under international humanitarian law or other rules of international law applicable to them". Final document of the Third CCW Review Conference, Part II, Geneva, 7-17 November 2006.

Complex Critical Systems: Can LAWS Be Fielded?

Martin Hagström*

Laws govern LAWS[1] – how to verify legality?

To be able to ensure lawful use of weapons, the user must understand the functionality of the weapon and how to operate it. The weapon must exhibit a predictable behaviour within the specified context of intended use, otherwise any attempt to take precautions against indiscriminate or disproportional effects will have an unpredictable outcome. Likewise, the functions of the weapon must be reliable in order to minimise the risk for a failure that can cause unintended effects. Information about weapons performance is for obvious reasons classified and non-sharable. In the current debate about so-called autonomous weapons, there is great concern about the predictability of such weapons. It is a challenge to formulate transparent requirements on weapons controllability and users' capabilities without disclosing a weapon's functionality and performance. This paper argues that information about the measures and procedures a state takes to ensure reliability and predictability of weapons might be less sensitive to share and might be a step forward in building confidence.

Weapons-critical systems

Systems, which, if they fail to perform as intended, can result in significant economic losses, physical damage, or threats to human life, are often labelled 'critical systems'. Systems with functions critical to safety exist in many domains and industries. Typically, such systems are found in the aerospace, nuclear power, rail, and weapons systems domains. Responsibilities and obligations for system safety are regulated in specific legislation. Legislation varies between domains, applications, and nations. For international areas such as civil aviation, the standards and regulations have to be multinational. In the case of civil aviation these standards are developed by the International Civil Aviation Organization (ICAO). The use of weapons is regulated both by international and national laws.

* Swedish Defence Research Agency.

1 Lethal Autonomous Weapons Systems.

Failure, or unintended effects, of a complex technical system is seldom the effect of one singular cause. A complex system is typically developed over a long period of time and by various actors. A complex system is often manufactured by several different actors and operated by an organisation different from the original producer. There are thus many possible reasons for undesired effects. To assign responsibility for failures can therefore be a difficult task. For this reason, detailed legislation, policies, and standards with the purpose to prescribe methods and development processes, with obligations for the developing and user organisation to follow exist in many domains. If a manufacturer follows the standards, structures the development work accordingly and keeps track of different technical contributions this will reduce the probability for errors to occur. In both military and civil domains, there are requirements regarding so-called 'catastrophic failure rates'. 'Catastrophic failures' denote failures which cause fatalities. They should generally be 'extremely improbable', often quantified as a probability of 1 or 10 in a billion.[2]

Critical systems can fail to perform as expected for several reasons: The hardware can break, the software can be erroneous, or the system can be handled in the wrong way. Hardware is tangible, measurable and in most cases testable. Hardware can fail for many reasons; shortcomings in design, manufacturing errors, fatigue, or general lack of maintenance. Systems whose functionality to a large extent depends on software can fail because the software may produce unexpected results or behave in an unintended way. A system may function as expected but the operator may fail to use it in the intended way, with undesired or even catastrophic consequences.

Operation of weapons systems

Unintended, negative outcomes are not only the consequence of technical failures. Operational failures are as important to prevent as technical failures. Users of weapons are required to understand how weapons should be operated to be able to handle them correctly. Users are also required to understand the effect of the weapons to be able to use them in a lawful way. Risk mitigation measures include careful design of man-machine interfaces, analysis of possible usages, and the development of doctrines, handbooks and training programs.

2 See e.g. Federal Aviation Administration, Advisory Circular 23.1309-1E, Figure 2, p. 23; also International Electrotechnical Commission Standard 61508.

Military forces have a long tradition of developing such tools. For obvious reasons, it is crucial to have well-trained soldiers and well-prepared commanders in warfare. Not only for the safety of personnel, but more importantly to achieve military goals. These goals are challenged if systems have unintended effects or if failures occur. Requirements for such doctrines, manuals and training programs are defined both in international documents such as NATO standards and in national regulations. On a national level, those requirements are often implemented in the form of detailed procedures and organisational structures.[3] The procedures prescribe and control that necessary operational guidelines and training standards are developed when a weapon is to be fielded.

With increasing system complexity, there will be an increasing need of analysis, planning, preparation, and training before a system can be fielded. The interplay between operators and machines will require more attention as system complexity increases. The research field of man-machine-interface is multi-disciplinary. There are numerous challenging research topics that need to be investigated when increasingly complex systems with progressively higher levels of automation are being developed. Future requirements on the handling and use of highly automated weapons systems need to be developed. Regulating bodies and military forces must follow the technological development to understand and plan for necessary improvement of methods and analysis techniques.

Critical systems

Development of technical systems might follow different processes depending on intended use and purpose. Different quality cultures grow in different markets and areas. For instance, manufacturers of consumer electronics only need to meet a quality standard proportionate to the price customers are willing to pay. There is no need for higher performance or quality than the market requires. There are no specific (legal) requirements regulating the performance of a robot vacuum cleaner, except in some safety-related areas like electricity (charging devices) and electromagnetic compatibility. The performance of a robot vacuum cleaner is not considered critical to anyone's immediate safety.

3 There are many documents available to the public. The U.S. Army has a list of field manuals, e.g. Unified Land Operations Army Doctrine Publication No. 3-0, Headquarters Department of the Army, Washington, DC, 10 October 2011; and the Swedish Armed Forces publish doctrines and handbooks, e.g. Handbok IKFN, Hävdande av vårt lands suveränitet och territoriella integritet, Swedish Armed Forces, 2016.

The cost, for consumers or manufacturers, for malfunctioning household appliances is limited and does not necessarily force developers to spend significant efforts on developing ultra-high quality products.

Conversely, the costs of failures that can have catastrophic consequences can be extremely high. In such cases high development costs are acceptable. Developing a system which can cause hazards and fatalities if it fails to function as intended thus increases focus on reliability.

This focus on reliability can be seen in the realm of safety of critical systems. Typically, safety standards are set by law – at the international and/or domestic level – to avoid hazards or dangerous situations. Depending on the kind of product, a standard can define different properties of the product. For example, it can be a specified strength (or material and size) for cableway cables, insulation of electrical equipment, construction products, etc.

For mechanical systems, the design specifications are in many cases concrete, measurable, and testable. Thickness of a steel cable or electrical insulation, protective shielding in engine rooms, and construction requirements for buildings are relevant cases in point. As technology develops, it can be expected that hardware eventually becomes too complex to test at a system level but mostly it is possible to actually test whether a product fulfils the requirements during the production phase.

For software systems it is not easy, if at all possible, to test a complete product. Software and computer programs are not tangible as physical objects. The execution of code is not observable, and only the output of the program can be measured. The use of computers in technical systems makes it possible to design systems with advanced, flexible capabilities. With advanced and flexible capabilities comes high complexity. Testing of complex systems with many 'components' is difficult since the number of possible cases that need to be tested quickly rises to a level where it becomes infeasible to do a complete functionality test at system level, i.e. when all components are put together to a complete system.

Software (the heart of autonomy)

Predictability and controllability are essential features of weapons systems that are to be considered legal according to international humanitarian law. This requirement is, in essence, not different from other complex safety-critical software systems. The software in the control system of an aircraft must be ultra-reliable. The regulating authorities of civil aviation, e.g. US FAA and European EASA, require that aircraft flying in civil airspace follow strict validation and verification programs during both design and operational phases. The regulations are

detailed and prescribe which design processes, verification, and validation methods should be used. Regarding the developer's organisational structure, regulations will typically require independence between test and development departments. Similar standards exist for rail systems and nuclear power plants. Recently, a standard for automation in the automotive industry has started to emerge. These standards have different classes of so-called 'safety integrity levels' to reflect the degree of severity of possible consequences of failures. If consequences of failures are minor in terms of hazards for humans or property, less rigid development processes can be employed. However, if the consequences could be catastrophic the strictest procedures must be followed in design, development, and use of the system.

Below is a table listing some standards that govern the development and use of safety-critical systems. It is notable that one of the first standards to emerge was the standard for military systems. One reason for this is of course that weapons are dangerous by design and there has always been a need for strict control of their development. With the introduction of computer control systems in civil aviation in the late 1980s came the need for a standard governing systems' development and use. The automotive industry published software guidelines already 20 years ago, although the need for a formal standard has become evident only recently when autonomous cars started to emerge from the laboratories.

Domain	Standard	Year	Levels of safety and consequences
Combat Systems	MIL-STD-882D	1977	IV (negligible) ... I (catastrophic)
Aircraft	FAA Do-178C	1992	Level D (minor) ... Level A (catastrophic)
Automotive	MISRA Software Guidelines	1994	SIL 1 (lowest) ... SIL 4 (highest)
Spacecraft	NASA NPG 8715.3	1996/1997	Negligible, Moderate, Critical, Catastrophic
Rail	CENELEC EN-50128/9	1997/1998	SIL 1 (lowest) ... SIL 4 (highest)
Medical	FDA	1998	Minor, Moderate, Major
Process Control	IEC 61508	1998	SIL 1 (lowest) ... SIL 4 (highest)
Automotive	ISO 26262	2011	ASIL-A (lowest) ... ASIL-D (highest)

All software standards have the purpose of increasing software reliability. Reliability does not come for free; products of high quality are, almost without exception, more expensive to develop than products of low quality. The estimates of increased costs for developing

software according to safety critical standards range from 25 percent to 100 percent.[4] This is, however, not necessarily an increased overall cost. When the use of the product depends on reliability, safety is cheaper than accidents. High quality software is likely to have a longer lifetime, and is cheaper to maintain and improve than low quality software products. Not only accidents and danger to human lives cause concern. There are numerous examples of extreme costs for software errors, like the one in a software deployed by the stock trading firm *Knight Capital.* The firm launched a new trading algorithm which autonomously bought shares on the stock market and then immediately sold them at a lower price. It was an unintentional behaviour and even though the difference in buying and selling price was mere cents, the loss was more than 460 million dollars in less than an hour, as there were thousands of transactions per second.[5] In a 2003 report[6] from the National Institute of Standards and Technology (NIST), cited in Computerworld,[7] the overall cost in the U.S. for software errors was estimated to 60 billion dollars annually at the beginning of the millennium.

Software errors, which cause the software to produce incorrect, unexpected results or behave in a way that is unforeseen and unintended, can come in many forms. There are many examples of strange and surprising effects as well as errors, which could have easily been avoided by using appropriate development procedures. Several of these examples have been collected and presented in different studies and overviews.[8] The flaws can stem from obvious programming errors which are detectable and reasonably easy to understand. However, unexpected behaviour can also arise from sheer complexity. In a large software system, there are many dependencies between different parts or modules of the software. It can be difficult to gain a complete overview and understanding of all dependencies. Therefore, the methods to develop safety-critical software prescribe thorough, detailed, and concrete development processes. Many costly failures could have been avoided if these

4 V, Hilderman, DO-178B Costs Versus Benefits, HighRely, 2009, DO-178B Compliance: Turn an Overhead Expense into a Competitive Advantage, IBM Corporation Software Group, 2010.

5 M. Philips, Knight Shows How to Lose $440 Million in 30 Minutes, Bloomberg, 2 August 2012, http://www.bloomberg.com/news/articles/2012-08-02/knight-shows-how-to-lose-440-million-in-30-minutes; 'Knight Capital Group', Wikipedia, https://en.wikipedia.org/wiki/Knight_Capital_Group.

6 G. Tassey, The Economic Impacts of Inadequate Infrastructure for Software Testing, National Institute of Standards and Technology, May 2002, https://www.nist.gov/sites/default/files/documents/director/planning/report02-3.pdf.

7 P. Thibodeau, Buggy Software Costs Users, Vendors Nearly $60B Annually, Computerworld, 25 June 2002, http://www.computerworld.com/article/2575560/it-management/study--buggy-software-costs-users--vendors-nearly--60b-annually.html.

8 G. Tan, A Collection of Well-Known Software Failures, 26 August 2016, http://www.cse.psu.edu/~gxt29/bug/softwarebug.html#economiccost.

procedures had been followed. The reason for the limited number of catastrophic failures due to software errors in safety critical systems, compared to failures due to software errors in other types of systems, is the strict use of standardised development procedures.

One way to keep a low error probability is to make the software simple, clear and not too long. This is, however, difficult when advanced complex systems are to be developed. It is still a challenge to develop complex safety-critical code, and such projects might suffer from delays and increased costs. One example is the US fighter aircraft program F35 which, with 8 million lines of code, has the largest software system in any aircraft. With increasing software size comes increased complexity. But complexity, and thus the development effort, increases faster than the software size itself. The difficulty to develop complex safety-critical software at an affordable cost is a great challenge. Large efforts are put into the development of methods and techniques to reduce these costs. The automotive industry is likely to be a driving force in this development as autonomous cars are expected to mix in with the existing, manually driven, vehicle fleet in the future.[9] However, whether expectations of a fast introduction of autonomous, self-driving cars will come true remains to be seen.

In many domains, trust and confidence about safety and predictability in critical systems are built by agreement on development standards. Industries can follow these development procedures and produce safety-certified products, without sharing proprietary technology. This enables competitive development of products with a shared confidence in quality. A state that uses weapons needs to make certain that the use follows relevant international laws. This requires system predictability and reliability, which, if software functions are involved, requires that strict development procedures are followed.

There are several obvious reasons for States not to share information about weapon systems performance. International non-proliferation agreements like the *Wassenaar Arrangement* on export control and the *Missile Technology Control Regime* prohibit the exposure of weapon technology. In several domains it is clearly possibly to form standards and share information about development procedures without revealing information about critical technologies. Thus the sharing of standardised development procedures in order to comply with international law does not mean sharing information either about performance of weapons or critical technology. This could mean a future path of trust and confidence for States to follow.

9 See e.g. C. Bergenhem et al., How to Reach Complete Safety Requirement Refinement for Autonomous Vehicles, CARS 2015 – Critical Automotive Applications: Robustness & Safety, Paris 2015.

Accountability for Lethal Autonomous Weapons Systems under International Humanitarian Law

Cecilie Hellestveit*

Introduction

Lethal autonomous weapons systems (LAWS) have two characteristics that challenge traditional rules of the international law of armed conflict, in the following referred to as international humanitarian law (IHL). LAWS expand the distance in time and space between the soldier and his military target by introducing a weapon that has its own agency. This allows for dislocating the link between the two sides to the point where it may be pertinent to suggest that LAWS is *replacing* the soldier in the hostile interaction. A number of questions concerning accountability under IHL consequently arise, including the question of whether machines *can* replace soldiers in the chain of command without distorting the entire system of IHL.

Rules of accountability under IHL refer to hierarchical command structures inherently required by IHL, rules of individual responsibility, and of state accountability. These rules serve to strengthen compliance and enforce rules of IHL by identifying who is accountable in case IHL is violated. IHL is distinguished from other legal frameworks applicable to the use of force because it governs and restricts an activity that by nature is *reciprocal*. IHL may only apply if there are two (or more) parties to an armed conflict that engage in mutual exchanges of hostile acts.[1] These norms regulate the encounter between two military hierarchies and adjacent structures. Often, this encounter takes place in the form of a physical meeting between two enemy fighters. The question posed here is what happens to the norms of IHL and its system of accountability when man fights with machines or against machines that have agency, in the sense of a capacity to make choices and act without human involvement in the hostile encounter.

The ability of a robotic system to address many questions and evaluations that are necessary for complying with IHL would require a level of programming and technical complexity beyond current technology, and perhaps even beyond that envisaged for the near future. This chapter does not evaluate the probability of such an evolution but instead focuses

* International Law and Policy Institute, Oslo; Atlantic Council, Washington DC.

1 See Common Art. 2 and 3 to the Geneva Conventions I-IV of 12 August 1949.

on the nature of IHL, the effects that introduction of LAWS is likely to have for rules of accountability in IHL in specific scenarios, and how LAWS are likely to challenge and potentially distort the existing regime of accountability under IHL.

For the sake of simplicity, this analysis will be limited to the use of autonomous weapons in conflicts that arise between states, and that consequently fall under the scope of application of Common Article 2 to the four Geneva Conventions.[2] Additional Protocol I to the Geneva Conventions (API)[3] applies to such conflicts, either by way of ratification, or by way of custom in as far as the main provisions relevant to LAWS expressed in API are declaratory of customary international law.[4] These principles apply irrespective of the means or methods of war utilised, and consequently apply to any use of LAWS in offence or defence in the course of an armed conflict.[5] The separate problems that may arise if LAWS are applied in law enforcement operations will not be addressed.

Accountability for violations of IHL – a delicate system

Rules of accountability under IHL have two distinct purposes. Firstly, they distribute responsibility for acts of war *within* the party to the armed conflict (1). Secondly, they distribute responsibility for certain effects of acts of war *between* the two adversary parties (2).

(1) The rules of accountability presuppose that a military organisation exists on each side that is responsible for the conduct of the party. IHL does not regulate anarchic violence, but is dependent on a certain level of hierarchy within each party.[6] A chain of command is a prerequisite for the very application of the rules of IHL. While this is made explicit for application of IHL in non-international armed conflicts,[7] it is a requirement inherent in Common Article 2's reference to a High Contracting Party, i.e. a state. It is the responsibility

2 The provisions apply to "all cases of declared war or of any other armed conflict which may arise between two or more of the High Contracting Parties".

3 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflict, adopted on 8 June 1977, entered into force on 7 December 1978 (API).

4 ICRC, Customary International Humanitarian Law, Volume I: Rules, Cambridge 2005; U.S. Department of State, 3 Cumulative Digest of United States Practice in International Law, 1981-1988, paras. 3434-35.

5 Art. 49 API.

6 Art. 80(2) API: each party "shall give orders and instructions to ensure observance [of IHL] and shall supervise their execution".

7 Art. 1(1) of the Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflict, adopted on 8 June 1977, entered into force on 7 December 1978 (APII): "organized armed groups [...] under responsible command".

of a party to have a system that enables compliance with IHL.[8] The system is outlined in Article 43 API. The Protocol stipulates a chain of accountability for compliance with IHL within the party along the lines of command. A military organisation enables a *chain of command* going from the top to the bottom, and a corresponding *chain of accountability* going from rank and file to the military commander.

Under a traditional command system, a broad array of enforcement mechanisms intends to prevent IHL violations, ranging from disciplinary sanctions by the party for soldiers disobeying orders,[9] via punitive measures imposed by the party for the behaviour of a soldier or a group of soldiers, to the prospect of criminal liability in case of capture by the enemy belligerent. If a soldier captured by the enemy has complied with IHL, he may not be prosecuted for his hostile acts on account of the privilege of having combatant status.[10] If the soldier has violated certain duties of IHL, however, such as the requirement not to blend in with civilians, he may nevertheless be prosecuted for his hostile acts.[11] If he has committed grave breaches of IHL, his privilege of immunity as a combatant does not protect him from prosecution for these crimes. He must stand trial for such violations by the belligerent adversary in case of capture, or by his own state upon return.

API is premised on a chain of command consisting of individuals with responsibility who can be identified directly or by way of command responsibility, and who can be held accountable for IHL violations and suffer individual sanctions.[12] While there is individual accountability for grave breaches of IHL, commanders are likewise responsible for breaches by their subordinates if they knew, or should have known, about relevant misconduct, and did not take all feasible measure within their power to prevent or repress the breach.[13] In cases where IHL violations amount to grave breaches of the Geneva Conventions and API, the party will be obliged to either prosecute or extradite those responsible to other countries or institutions in

8 Art. 86(1) API: the party shall "require military commanders, with respect to members of the armed force under their command and other persons under their control" to "prevent [...] and suppress" breaches of IHL.

9 Art. 87(3) API.

10 R. Baxter, So-Called 'Unprivileged Belligerency': Spies, Guerrillas and Saboteurs, 28 British Yearbook of International Law 323 (1951).

11 Art. 44 and 46 API.

12 Art. 49 Geneva Convention I; Art. 50 Geneva Convention II; Art. 129 Geneva Convention III; Art. 146 Geneva Convention IV.

13 Art. 86(2) API; under customary law, the *mens rea* element is that the commander "knew or had reason to know" that subordinates were about to commit or were committing a war crime. See ICRC (n 4), Rules 152 and 153. Under the Rome Statute of the ICC, command responsibility covers situations where the commander "either knew or, owing to the circumstances at the time, should have known" that forces were about to commit or were committing war crimes; see Article 28(a)(i) of the Rome Statute of the International Criminal Court, adopted on 17 July 1998, entered into force on 1 July 2002.

order to ensure prosecution. Identifying accountability for acts linked to the armed conflict within this chain of command is supposed to be an *intuitive* undertaking. Every solder must be aware of his obligations under IHL. This is supposed to reduce the scale and nature of IHL violations.[14] Correspondence between the chain of command and the chain of accountability is essential for the system of IHL to function properly.

LAWS may have certain qualities that resemble those of a human soldier. However, LAWS cannot be held accountable in the sense of IHL.[15] Consequently, the following question arises: *who will assume accountability in their place*? If LAWS are the agent, which soldier will have to assume responsibility for the acts of LAWS *within* the party to the armed conflict? LAWS are complicated systems, where numerous persons and entities are responsible for proper functions and effects – some of which are outside of the traditional chain of command. While this trespassing of the chain of command is a common feature of modern warfare, when associated with LAWS it exposes the accountability regime of IHL to even more pressure, since the entity equipped with agency on the battlefield *cannot* be held accountable.

(2) In the encounter between two enemy parties to an armed conflict, IHL establishes certain principles and rules that apply *reciprocally*, meaning that provisions of IHL apply *with equal force to both parties*. This extends to the principle of distinction, the rule on proportionality, the duty to take precautions in attack, and the prohibition to cause unnecessary suffering or superfluous injury. The obligations are accompanied by corresponding duties of precaution in defence.[16] These principles apply irrespective of the behaviour of the adversary, since humanitarian rules are not subject to reciprocity as enforcement mechanism.[17] When the adversary party is in breach of its obligations under the rules of conduct of hostilities, this may alter certain rules of accountability.

With respect to rules of conduct of hostilities, in some very specific circumstances, such as immunising military objectives by using involuntary human shields, the use of child soldiers, or the feigning of civilian immunity, accountability for the unlawful effects will

14 Art. 87(2) API.

15 K. Egeland, Lethal Autonomous Weapons Systems under International Humanitarian Law, 85 Nordic Journal of International Law 89 (2016).

16 Art. 58 API.

17 Art. 60(5) of the Vienna Convention on the Law of Treaties, 1969, entered into force 27 January 1980, precludes a state from suspending or terminating for material breach any treaty provision "relating to the protection of the human person contained in treaties of a humanitarian character". See also Article 51(8) API, recalling that indiscriminate attacks do not release the Parties from their obligations. Belligerent reprisals are still permissible, but subject to stringent restrictions. They are proscribed against protected persons.

not rest with the party using force, but will be placed on the adversary party. The objective of this distribution of accountability between the parties is to protect the rules of IHL from being undermined by parties exploiting protection under IHL in order to gain military advantages that would undermine respect for IHL itself, to the detriment of all. This 'shift' of accountability from one party to the other is a part of the internal enforcement mechanisms of IHL, and a way to prevent abuse and the weakening of IHL protections over time.

Rules of accountability for violations of IHL are hence not merely a set of singular rules put together by chance, but constitute a rather delicate system that distributes accountability *within* a party to an armed conflict and *between* the parties to the armed conflict for the purpose of enhancing and preserving respect for these rules to the benefit of all. If LAWS perform the tasks of soldiers, how will this affect the distribution of accountability *between* the parties?

In the following, a closer look is taken at how LAWS must be expected to influence, distort, and affect this system of accountability for IHL violations. Three archetypical violations of IHL by LAWS are likely to present different problems. LAWS may violate IHL by intention of the party, by design, or by accident. The main emphasis will be on the last point, where the major difficulties linked to LAWS are expected to arise.

Violation by intention: LAWS violate IHL as a deliberate effect

A party may use LAWS to conduct hostilities in a way that intentionally contradicts or ignores IHL. In such a scenario, most rules of accountability for IHL violations will apply in ordinary ways.

LAWS may be intentionally programmed to directly target civilians, civilian objects, or persons *hors de combat,* in which case they will violate the principle of distinction.[18] LAWS may alternatively violate the prohibition on indiscriminate attacks if they are programmed in ways that strike military objectives and civilians or civilian objects without distinction, either because LAWS do not or cannot direct attacks at specific military objectives,[19] or because LAWS have effects of a nature that is indiscriminate.[20]

18 Art. 51(2) and 52(1) API; Common Art. 3(1) GC.
19 Art. 51(4)(a) and (b) API.
20 Art. 51(4)(c) API.

The duty of precaution in attack extends responsibility for IHL violations to those who plan or decide upon an attack.[21] It imposes duties on the soldier launching the LAWS, and the commander, according to traditional rules of accountability under IHL as described above. This duty of precaution may also extend to the programmer, who does not plan or decide upon an attack, but whose contribution determines the agency of the LAWS.

In such situations, therefore, the ordinary rules of accountability for IHL violations will apply to the use of LAWS. These violations may amount to grave breaches of the Geneva Conventions,[22] and are regarded as war crimes.[23] This entails responsibility by the Party for the violation (i.e. state responsibility),[24] individual responsibility for those responsible for preparing and programming the LAWS in ways that would violate or ignore the rules,[25] and possibly command responsibility for the military commander.[26] Consequently, commanders and civilian supervisors can be held accountable for these war crimes if he or she "knew or should have known" that the autonomous weapons system had been programmed in such a way.[27] Accountability, in the sense of holding states and individuals responsible, when a party *intentionally* violates or disregards IHL by using LAWS is not likely to pose problems out of the ordinary. In this sense, LAWS are a weapon like any other.

Violation by design: LAWS lack the ability to comply with certain aspects of IHL

LAWS may lack the ability to comply with IHL. Accountability under IHL applies to LAWS on an equal footing with other means of war. LAWS are consequently subject to review under Article 36 API for the purpose of determining whether the employment of a given type of LAWS would, in some or in all circumstances, be prohibited under IHL applicable to the party. This is an obligation that applies to State Parties to the API; its status as a norm of customary international law is currently unclear. The review must assess whether a specific type of LAWS is designed, manufactured, or programmed in a way that may cause effects that are unlawful under international humanitarian law, more specifically whether the

21 Art. 57(2)(a) API.
22 Art. 85(3)(a)-(f) API.
23 Art. 85(5) API; see also the corresponding rules under the Rome Statute for the ICC. It is perceived to be declaratory of custom that violation of these rules in international armed conflicts will amount to war crimes.
24 See International Law Commission, Draft Articles on State Responsibility for Wrongful Acts; Art. 91 API.
25 Art. 87(3) API.
26 Art. 86(2) API.
27 M.N. Schmitt, Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics, Harvard National Security Journal Features, 2013, http://harvardnsj.org/wp-content/uploads/2013/02/Schmitt-Autonomous-Weapon-Systems-and-IHL-Final.pdf, p. 33.

weapon can and will comply with the principle of distinction (i.e. the duty to only target military objectives, prohibition against indiscriminate attack, including the rule of proportionality), the prohibition against unnecessary suffering or superfluous injury,[28] and asses its effect in terms of qualified damage to the natural environment.[29]

Review of weapons is compulsory for States Parties to the API. This entails an obligation to assess whether the weapon under ordinary circumstances and use would cause unlawful effects. LAWS may pose problems on account of technical complexities and on account of the level of autonomous agency. These weapons will therefore require extended review, which is likely to include a determination about the level of human control necessary over the weapon, the ability to abort an attack, and reassuring methods to enable a soldier or a military commander to prevent breaches of IHL.[30] The higher the level of autonomy, the more extensive the review and the more reassuring guarantees are likely to be necessary under Article 36 API.

Violation by accident: LAWS incidentally violate IHL

LAWS's ability to have agency and to act with a certain level of autonomy introduces a range of new ways in which a weapon may violate IHL, not by intention or design, but by varying levels of 'accident', in the sense of autonomous agency. These are the situations where LAWS raise the most intriguing and complex problems linked to accountability for IHL violations.

The soldier has autonomy of agency. He has his place in the chain of command with corresponding rights and duties. Rules of accountability under IHL for violations mirror this system. If LAWS increasingly perform some of the tasks commonly executed by the soldier, who will be accountable for the violations of IHL that the LAWS may incidentally commit? If LAWS incidentally direct an attack at civilians in violation of the principle of distinction,[31] or fail to recognise surrender by a combatant *hors de combat*,[32] who is accountable for this violation of IHL? As LAWS cannot themselves be held accountable in the sense presumed under API, who will be held accountable in their stead?

28 Art. 86(2) API.
29 Art. 35(3) API.
30 Art. 87(1) API.
31 Art. 51(2) API.
32 Art. 41(1) API.

Three hostile encounters are reviewed here. The encounter between LAWS and civilians, the encounter between LAWS and enemy combatants, and the hostile encounter between two LAWS.

LAWS vs. Civilians

The rules on the conduct of hostilities most likely to be exposed to incidental violation in an encounter between LAWS and civilians are those linked to the principle of distinction. This includes the duty of a party to select targets that may be lawfully engaged, and only direct attacks at such targets.[33] If an attack is directed at a lawful target, it must not be expected to cause incidental harm to civilians and civilian objects which is excessive in relation to the direct and concrete military advantage anticipated (i.e. the rule of proportionality).[34] Finally, a party must take all feasible precautions during the entire targeting cycle with the aim of sparing the lives of civilians (i.e. the rule of precautions in attack).[35]

LAWS may incidentally directly target civilians or civilian objects. Alternatively, LAWS may cause disproportionate civilian casualties. The duties to avoid such effects under IHL are framed as *ex-ante* evaluations, prior to and during the targeting cycle. These evaluations are context-specific, and, from the perspective of a machine, very complex. Can LAWS make such evaluations in a meaningful way at all?

The accountability chain of a party under IHL implies that a combatant has individual accountability, while his superiors have command responsibility. The entire system under API is premised on the possibility of identifying individuals who are responsible for the various acts of the party. How will this play out if LAWS have autonomy in the sense of agency?

Firstly, it may be difficult to identify the failure, and consequently to determine which entity bears responsibility for such violations of IHL. When the cause is neither intent nor design, the process of determining who is responsible in place of LAWS for a given violation is likely to be a complex process. Accountability in IHL, in contrast, is expected to be immediate and intuitive in order to have the desired effects of inducing compliance with IHL.

33 Art. 51(2) API.
34 Art. 51(5)(b) API.
35 Art. 57 API.

Secondly, identifying who is accountable is further complicated by the high level of civilian technology involved with LAWS. This distributes the potential for accountability beyond the military chain of command, to a manufacturer or programmer, distorting the correspondence between the chain of command and the chain of accountability in API.

Finally, the large number of individuals and entities involved in the design, manufacture, programming, preparation, and dispatching of LAWS further pulverises individual accountability as envisaged by API. While this is a common feature of manufacture and operation of modern weaponry, the absence of human agency in the hostile encounter when LAWS are involved increases the detrimental effects of pulverisation for questions of accountability.

The way in which a party to a conflict may take measures to prevent such violations by default is by having reassuring procedures of precaution. It is therefore likely that LAWS will further increase the emphasis on duties linked to precautionary measures – in particular the duty to do everything feasible to verify that LAWS are able to distinguish lawful military objectives from persons and objects that enjoy immunity from attack,[36] and the duty to take all feasible precautions in the choice of means to avoid and minimise indiscriminate effects,[37] including the obligation to choose a different weapon in circumstances when violations of IHL might occur. A major hurdle is that precaution in attack is presently among the opaquest elements of IHL when it comes to accountability for conduct of hostilities.

While many of these violations of IHL are likely *not* to qualify as grave breaches of IHL and war crimes, the distortive effects that LAWS are likely to have on the internal enforcement mechanisms of IHL in international armed conflicts should not be ignored or taken lightly. These dynamics may cause LAWS to have long-time detrimental effects on enforcement of IHL far beyond operations involving LAWS, with serious and potentially devastating effects for the respect for IHL more generally. The current accountability model under IHL is based largely on a logical and intuitive function in the chain of command, mirrored in the chain of accountability. This system will be jeopardised by LAWS, raising questions about the need for a separate system of accountability for operations involving LAWS.

36 Art. 57(2)(a)(i) API.
37 Art. 57(2)(a)(ii) API.

LAWS vs. Combatants

Problems of a different nature arise when LAWS meet enemy combatants in a hostile encounter. The restrictions that apply to both parties, and by extension their agents, is the prohibition to target a combatant who has been placed *hors de combat* by injury or has given a clearly expressed indication of the intention to surrender.[38] The first question that arises is whether LAWS must be able to identify an enemy combatant who is signalling surrender. If not, LAWS will effectively undermine one of the most important pillars of IHL, namely the right to quarter and prohibition to attack a defenceless combatant – a cornerstone in the rules of IHL that aim to offer combatants an avenue of exit in battle other than the choice of killing or being killed. This is seen to have an important humanising effect, influencing how soldiers approach battle and conduct hostilities. The supposition is therefore that LAWS must be able to identify a soldier clearly expressing indication of intention to surrender.[39]

The next question that arises is whether the combatant has corresponding duties. In classical hostilities with human encounters, the combatant signalling intention to surrender will be under the obligation not to feign immunity and abuse the protection of IHL in order to gain a military advantage.[40] The permission of ruses of war and the prohibition of perfidy are constitutive elements of a delicate system of enforcement of IHL that distributes accountability *between the two parties* for the purpose of strengthening respect and enforcement of the rules of IHL. The basic idea is that there are humans on both sides, and the dynamic between humans must be such that even in the midst of hostilities, a fair balance remains, so *humanity* is preserved.

When one side replaces its frontline agents with LAWS, does this imply that the balance shifts? The party employing LAWS is under an obligation to ensure that LAWS do not violate prohibitions under IHL, including the duty not to target an enemy soldier who is incapacitated by injury.[41] The next question is whether there will be a corresponding obligation on the combatant not to feign incapacitation in the encounter with LAWS. Is man only allowed to fool the machine in ways that can be considered 'fair'?

38 Art. 41(1) API.
39 Art. 41(2)(b) API.
40 Art. 37(1)(b) API.
41 This is a grave breach amounting to a war crime, see Art. 11(4) API.

If the answer is yes, because the mentioned rationale – that taking advantage of the rules will weaken IHL more generally – applies with equal strength, does this mean that soldiers will be obliged to comply with the rules "as if the LAWS were a combatant"? This would entail that the LAWS takes the place of a combatant in terms of rights and privileges under IHL. And can it be justified that the soldier is expected to treat the LAWS with the same amount of fairness as if the machine were a soldier – for example by observing a corresponding duty on the soldier not to attack the LAWS if it becomes incapacitated. Is it even feasible for the internal enforcement mechanism of IHL of distributing accountability between the parties to function as long as LAWS cannot be held individually accountable? The soldier can be charged with perfidy and may be executed. The LAWS will not face a similar fate.

If the answer is no, because the rules of ruses and perfidy were made for honourable battles between human soldiers, and that a machine cannot be afforded the same privilege, the next question is – what follows from that? Will the soldier be allowed to resort to any measure to *fool* the autonomous weapon? Will the autonomous weapon consequently be forced to adapt to this tactic, pushing programming into a 'presumption of perfidy by humans'? And how does this influence the level of lawful use of force employed by the machine? How will such a scenario play out for the combatant in the end, and will this not inevitably introduce a sub-set of rules applicable to encounters between human and machine that is different from that for the encounter between human soldiers?

If ordinary rules apply, the combatant is likely to end up at the losing end. Forcing the same rules for human and machine will provide the machine with a comparative advantage, sliding armed conflict between states into patterns of ever stronger asymmetry – always to the detriment of militarily inferior states – alas *without* high human costs for major military powers. It is also likely that this development will affect distribution of accountability between the parties, in the sense that the party relying on humans will end up as the violator of IHL. Over time, the rules on conduct of hostilities in interstate conflicts will increasingly face the dilemmas and quagmires associated with asymmetric, irregular warfare – ultimately to the detriment of all.

If ordinary rules do not apply, this would entail the introduction of a sub-set of norms for LAWS.

LAWS vs. LAWS

The introduction of a sub-set of norms becomes even more palpable in the event of an encounter between adversary LAWS. Presumably, such a hostile encounter will require programming that in part or in total will exclude humans from the fight. IHL is a system of norms that channels the risks associated with armed conflict in certain directions. The main function is to direct violence away from civilians by allowing for combatants to target and be targeted. When LAWS meet LAWS, humans are in principle out of the equation. Is it then pertinent that the principle of distinction extends to all humans, or that the combination of the prohibition to target civilians be joined by the prohibition of unnecessary suffering or superfluous injury, to construct a ban for machines to directly target *humans* in such battles? If LAWS nevertheless cause unlawful effects, the question of accountability is likely to open a new battlefield – which party is responsible for incidental effects of a battle between their respective machines? The type of accountability for violations of IHL under such a scheme will be very different from IHL accountability as we know it.

This development may effectively result in the *de facto* development of three sub-sets of rules for conduct of hostilities: one for battle between human combatants, a different set for battles when humans meet machines, and a third subset for encounters between machines. It would require a substantially more detailed (and problematic) regulation of hostilities than what is presently the case under international law.

Concluding remarks

The introduction of LAWS will trigger difficult questions of accountability under IHL. In cases where LAWS are used to commit deliberate violations of IHL, the system of accountability under IHL will work in ways comparable with war crimes committed with other weapons. Ultimately, this is due to the fact that an *intention* may be traced back to one or more humans in charge of the LAWS at some point in time.

The introduction of LAWS as a means of warfare is nevertheless likely to challenge fundamental lines of accountability under IHL. The range of challenges to accountability outlined above are linked to the notion that no identified human within the chain of command is directly accountable for the agency and autonomy of LAWS – causing a cascade of derived effects for how accountability for the rules of conduct of hostilities of IHL is distributed and how it will work out on the battlefield. LAWS will affect the distribution of accountability for violations of IHL *within* a party to an armed conflict. LAWS will also disturb distribution of accountability for violations of IHL *between* the two enemy parties

to an armed conflict. It is submitted that these disturbances are likely to affect the regime of IHL in profound and distortive ways. Over time these dynamics will most probably destroy the underlying, fine-tuned balances and internal enforcement mechanisms of IHL applicable to hostilities in armed conflicts between states – to the detriment of all humans.

A Human Rights Perspective on Autonomous Weapons in Armed Conflict: The Rights to Life and Dignity

Christof Heyns*

Introduction

Traditionally, users of weapons have been in direct physical control of their weapons. Over the years, revolutions in military affairs have produced weapons with increased range and fire-power, but, by and large, this did not change the fact that the person activating or launching the weapon took the decision when and against whom force would be used while being present on the spot from where force was projected.

The advent of unmanned or human-replacing systems has meant that the person launching the weapon no longer needs to be physically present at the place and time when it is released. The first generation of these unmanned systems are remote-controlled weapons systems. The best-known iteration of this technology is armed drones.

The emphasis here is on the next generation of unmanned systems in armed conflict. So-called 'autonomous weapons' (sometimes simply called robots or machines, or 'killer robots') would allow for the release of force from unmanned systems that is no longer remotely controlled by humans. Instead, once a human has activated an autonomous weapons system, on-board computers will make the determination, independent from direct human intervention, against whom it should be directed and when to release force.

The 'autonomy' of robots is not comparable to the autonomy of human beings, which is often seen as the basis of the ability of humans to act as free moral agents. However, robots with a high level of autonomy can perform functions that those who programme or deploy them cannot foresee (and in that sense, loosely speaking, autonomous weapons have something that might resemble 'free will'). 'Automatic' systems respond in predictable ways to their environment. 'Autonomous' systems may sometimes respond in unpredictable ways, and are, to the extent that this happens, outside human control. This is because machine

* Professor of Human Rights Law, University of Pretoria; United Nations Special Rapporteur on extrajudicial, summary or arbitrary executions (2012-2016).

learning takes place, which means that they alter their behaviour based on their experience. Moreover, especially in the context of something as chaotic as war, not all scenarios can be foreseen and provided for. As such, machine autonomy, beyond a certain point, can potentially undermine or limit human autonomy and control over the world.

The emergence of autonomous weapons will make it possible for humans not only to be physically absent from the point of release of force, as with armed drones, but to be 'psychologically' absent as well, to the extent that they do not take the on-the-spot decision to direct and open fire.

According to philosopher Nick Bostrom, the appropriate response of human beings to the advent of artificial intelligence in all walks of life is the central question of the day. An important perspective from which to view autonomous weapons would be the one of human rights. Human rights are widely accepted as the dominant normative – ethical and legal – framework of the international community. Indeed, Louis Henkin has famously called human rights "the idea of our time".

Clearly autonomous weapons can potentially affect the right to life, often described as the 'supreme right'. It could possibly also affect the right to dignity, a less clearly defined right but one that has played a foundational role in the development of both human rights law and international humanitarian law.

The potential advantages of autonomous weapons should be recognised. Unmanned systems in general offer the advantage of the protection of one's own forces – a significant protection of life advantage. Autonomous systems in particular offer clear additional potential military advantages to those who deploy them, such as increased reaction speed in decision-making and 'eyes' on the target, potentially increasing accuracy. Machines may also in some cases avoid mistakes made and even atrocities committed by humans on the battlefield because of emotions such as fear, fatigue, or revenge. It should be recognised that human compliance with international humanitarian law (IHL) is appallingly low, and if technology can improve that, the gains will be significant.

If considerations such as the above lead to better targeting, as proponents claim, autonomous weapons can save the lives of those who may not be targeted, such as civilians not engaged in the conflict. It is on that basis that roboticist Arkin and others have defended the development and use, under certain circumstances, of autonomous weapons. Clearly, neither from a military nor a human rights point of view can such potential capabilities be dismissed out of hand. To what extent is it possible to ensure that the benefits offered by autonomous weapons are used but its disadvantages are avoided?

I presented some of the questions raised by autonomous weapons as Special Rapporteur on extrajudicial, summary or arbitrary executions in a report to the United Nations Human Rights Council in 2013. At the time, I argued that states should impose national moratoria on the development and use of autonomous weapons until such time as an internationally acceptable way of dealing with increased autonomy in targeting has been found. In the meantime, the global engagement with this issue has advanced to the point where it may be easier to come to a firm view. This issue has been taken up in a number of international fora and has been subjected to thorough multi-disciplinary consideration.

There is a growing view that the dividing line between acceptable and not acceptable machine autonomy is determined by the question whether humans retain 'meaningful human control' over force delivery. Where meaningful human control is present, autonomous weapons are acceptable, but where there is no meaningful human control – where there is in other words 'full autonomy' on the part of the machine – they should be banned. I share this view in terms of which all forms of autonomy in targeting are not rejected; it is only when a certain threshold is reached in terms of excluding human control that machine autonomy is considered unacceptable.

In order to ascertain more clearly how the two rights under consideration are affected by full machine autonomy, it is useful to articulate the two primary questions raised by autonomous weapons in the context of armed conflict as clearly as possible.

Can they do it? Autonomous weapons and proper targeting

Can autonomous weapons do proper targeting? That is, can they be deployed in a way that has results that comply with the rules IHL sets for humans as far as the conduct of hostilities is concerned, and, in particular, those rules that protect some people on the battlefield from direct targeting (such as uninvolved civilians and the wounded) and prohibits excessive incidental or collateral casualties? Autonomous weapons may strike the wrong target altogether, or cause excessive harm to those incidentally affected by the use of force. The concern here is thus with *those who are otherwise considered to be protected under IHL*.

The right to life continues to apply during armed conflict. As far as the conduct of hostilities is concerned, the question whether a depravation of life is 'arbitrary'" and thus unlawful under international law is determined with reference to the rules of IHL. To the extent that the rules of IHL are not followed when autonomous weapons deploy lethal

force, the right to life is violated. The right to life has two components: the prohibition on the 'arbitrary' taking of life, and accountability where that occurs. The right to life is not an absolute right but the limitations are narrowly construed.

The IHL rules of distinction, proportionality, and precaution in attack are of specific importance in this regard. The rule of *distinction* seeks to minimise the direct impact of armed conflict on civilians and others who are not (or are no longer) part of the hostilities, by prohibiting the targeting of such people as well as indiscriminate attacks.

The traditional approach is that there is no prohibition on engaging those who are not protected, provided all the rules of IHL are complied with. The rules of IHL must be complied with as a collective in each attack. Commanders are allowed to target categories of people based on status or conduct; they are not required to identify specific individuals who pose an immediate lethal threat as targets, as is the case with the use of deadly force under human rights law. In situations where autonomous weapons cannot reliably distinguish between, for example, combatants or other belligerents and civilians, their use would be unlawful.

There are several factors that could possibly impede the ability of autonomous weapons to operate according to the rule of distinction, including the technological inadequacy of existing sensors; a robot's inability to understand context; and the difficulty of translating IHL language and definitions of civilian and combatant into computer programming.

The rule of *proportionality*, for its part, requires that the expected harm to civilians (so-called incidental or collateral damage) be measured, prior to the attack (as a whole, not each individual engagement), against the anticipated military advantage to be gained from the operation on the basis of reasonableness. If the expected harm to civilians outweighs the military advantage, the attack should not proceed. Compliance with this complex rule of international humanitarian law is largely dependent on value judgments and subjective estimates of the risks, and it can be questioned to what extent robots may reliably make such assessments, even more than is the case with distinction.

Whether an attack complies with the rule of proportionality needs to be assessed on the merits of each individual case, depending on the specific context and considering the totality of the circumstances. The requirement of a case-by-case analysis seems to militate against the idea that a single algorithm can be made available in advance to provide all the answers, and rather call for human case-by-case discretion.

It appears from the above that autonomous weapons may be able to meet the requirement of distinction only in limited cases, and even less as far as proportionality is concerned. The most likely cases where they will meet the requirements of IHL arise if autonomous weapons are deployed in an area where there are no civilians present. However, this is hardly an example of them making adequate distinction and proportionality assessments. Rather, in such cases, the distinction and proportionality decisions have been taken by the human operator who chose the terrain where the robots are deployed, serving as an exercise of human control from the 'wider loop'. Using machines that employ facial recognition to identify a target also does not amount to the machine engaging in distinction – that was done by the programmer who identified the particular person as a target.

The problems outlined above – that the lack of human judgement may prevent autonomous weapons from taking appropriate targeting decisions – obviously increase as human control over such decisions decreases, suggesting that developing a notion such as 'meaningful human control' can help to set the cut-off line.

All of the above is subject to the consideration that targeting technology is certain to evolve over time, and may in the process diminish the need for human involvement in order to be accurate. Moreover, to the extent that technology can assist targeting, there may be a duty on commanders to use it.

Lastly, the question whether autonomous weapons are able to release force in a reliable way brings to the fore the further issue of the yardstick against which their performance is to be measured. Some authors have argued that to be acceptable, autonomous weapons will be required merely to match the performance of the average human soldier.

This may require further consideration. Higher levels of control imply higher levels of responsibility. In order to be considered potentially acceptable, autonomous weapons will have to withstand closer scrutiny than the average human soldier, even if the objective standards remain the same.

Should they do it? Robots and the power to determine who will live or die

Most commentators would say that only the first question needs to be asked. To the extent that robots can at least match human targeting, they should be allowed, and if they can surpass humans, even more so, and if that is not the case they should not be permitted. There is however also the principled question whether it is not inherently wrong – as a matter of morality and of law – to let machines decide who should live and who should die, even if they can do proper targeting.

While the discussion in the previous section dealt with the plight of civilians and others who are protected, the lens here is broadened also to cover *the deployment of autonomous weapons against those who are not protected*, and can thus be legitimate direct targets or incidental casualties. The focus is thus on the manner of targeting; it is through a fully autonomous robot. All else being equal, is it of ethical or legal importance whether a human or an algorithm makes the determination that someone has to die?

One often hears the argument that it makes no difference who or what released a missile that hit someone – the result is the same, namely death, as long as the requirements of IHL are otherwise complied with. There is especially little sympathy in this context for those who may be directly targeted (e.g. enemy combatants), but the notion that the manner of targeting does not matter extends also to those who are incidental casualties.

I want to argue that there are right to life as well as right to dignity reasons why, where the fate of those who are not otherwise protected by IHL are at stake, full machine autonomy over critical functions should not be permitted.

The right to life is typically framed as a right against the 'arbitrary' deprivation of life, and an 'arbitrary' deprivation of life can be seen as a taking of life that is in violation of international law. Indiscriminate targeting or excessive collateral damage, as discussed above, constitute a violation of the right to life. Is autonomous targeting inherently 'arbitrary'?

Philosopher Peter Asaro has argued that an implicit requirement found in IHL is that lethal force may only be used based on a human decision, which cannot be delegated to an autonomous machine. One is not allowed to delegate a power that one does not have – and humans do not have the authority to use force without applying their minds. Non-human decision-making regarding the use of lethal force is, according to this argument, inherently 'arbitrary', and deaths that result are unlawful deprivations of life.

Many of our actions seem indeed to be premised on the idea that, where people's lives or other grave consequences are at stake, a human – someone identifiable – should apply their mind and assume ultimate responsibility.

For example, the idea that computers should serve as judges in courts of law does not seem to attract many serious supporters today, in spite of all the advances in technology. This is the case especially as far as criminal law is concerned, given its far-reaching impact on the human person. The notion that computers can assume final 'responsibility' to impose the death penalty or even sentences involving deprivation of freedom is unthinkable, even in the most authoritarian regime.

One way of interpreting the prohibition on the arbitrary deprivation of life is to argue that the involvement of computers in targeting decisions, given their quantitative approach, is the surest way to avoid arbitrariness. It offers an objective as opposed to subjective way of determining how to direct force.

However, as we saw earlier, autonomous weapons are to some extent unpredictable. While computers may be able to respond in a consistent way when presented with the same stimuli, the complex environment of war seldom presents itself twice in exactly the same detail.

The decision to use force against a human being, even if done in the heat of battle, requires at some level – somewhere in the decision-making loop, not necessarily on the front line – what Asaro calls an element of deliberation. Someone must take the time to apply his or her mind and 'sign off' on the conclusion that using force is really necessary as far as possible *in this particular case*.

Clearly, war already involves mass, long-distance killing, which entails a significant departure from the careful consideration of the merits of each use of force. However, the introduction of autonomous weapons, seen in the context of the exponential and seemingly unstoppable growth of technology, brings with it the potential of a wholesale abandoning of the ideal that the life of an individual should be taken only if it is the last available alternative to protect life in the situation in question.

Developing autonomous weapons involves taking the decision to kill people in advance, in the abstract, and based on hypothetical scenarios. Even more so, when a deadly algorithm is developed, or a fully autonomous weapon activated, the person who does this can by definition not have the full picture in mind which will apply when force is eventually released.

However, the right to life is violated not only when a particular deprivation of life is arbitrary, but also when there is no or inadequate accountability. The absence of accountability, in itself, also constitutes a violation. In the context of armed conflict, there must be accountability where the rules of IHL or international criminal law have been broken. Even if it often does not happen in practice, it is important that it is possible in principle.

If autonomous weapons outperform humans as far as targeting is concerned, there will be fewer mistakes, but there will still be mistakes. The question is who will assume legal and moral responsibility for those mistakes. It has been argued that autonomous weapons leave an accountability gap, meaning that from this perspective they also violate the right to life.[1]

To the above consideration of the right to life, an exploration of the impact that autonomous weapons furthermore have on the right to dignity should be added. The right to life and the right to dignity cannot be divorced – they are, in the standard language, interrelated and indivisible. What is at stake is the protection not merely of 'bare life', or the continuation of biological existence, but the protection of dignified life. Even if there is a legitimate limitation on someone's right to life, as is the case where the person is not protected under IHL, that person retains his or her other rights, such as the right to dignity.

It seems worth exploring more specifically what the dignity component of the right to life brings to the discussion in the context of autonomous weapons, again focusing in particular on the use of autonomous weapons against those who may otherwise be targeted.

Human rights in general, and dignity in particular, emphasise the notion that each person is entitled to be treated according to his or her own full merits. In the words of Ronald Dworkin, the concept of human rights requires that each person is entitled to "equal concern and respect". This means we should not simply be treated equally but also need to be taken seriously as separate and irreplaceable individuals.

This notion is directly challenged by the idea that people, including opponents in war, can become a casualty of the algorithmic calculations (literally reducing its targets to the 0's and 1's of the digital code) of an unthinking entity if we happen to be in its way. Where that occurs, one person's death is indistinguishable from that of so many others who happen to find themselves in the striking range of generic killing machines. Restrictions on the use of force are not simply a numbers game – it is also about protecting the value of each individual life.

1 Much has been written about the topic, which is why I will not go into this in more detail.

Closely associated with the notion of human dignity is the idea that humans should not be treated as something similar to an object that simply has an instrumental value (as is the case e.g. with slavery or rape) or no value at all (as with many massacres). The person against whom the force is directed by autonomous weapons is reduced to being an object that has to be destroyed, and that is even more clearly the case where incidental casualties are at stake. They have no avenue, futile or not, of appealing to the humanity of the enemy, or hoping it will play a role, because there is a machine on the other side.

The realities of modern warfare are of course such that in many cases, someone who is about to be targeted does not have a real chance of appealing to the humanity of the person on the other side. However, the hope that this may be possible has so far not been completely excluded. With the introduction of autonomous weapons there is no such chance. Having autonomous weapons as a legal and legitimate part of the world which we live in can undermine an important part of our hard-wiring: namely, hope.

Dignity can be approached from different angles. There is the dignity of the person at the receiving end, under discussion above, who is reduced to being a target or incidental damage. But the autonomous use of force can also implicate the dignity of the people on the side that deploys autonomous weapons, because it is being done in their name. Dignity entails the ability to be a moral agent: to be autonomous in the sense that they exercise moral choices and to be responsible for their outcomes. Fully autonomous weapons as a human-replacing technology may remove this possibility to be a moral person. Conversely, where there is indeed meaningful human control, autonomous weapons can potentially increase human autonomy and control over the outcomes of one's actions, and thus enhance human dignity.

Because the concern of the "should they?" question is with the fate of those who are not protected by IHL from attack – issues of legal accountability for the possible violations of the right to life and the right to dignity do not arise in this part of the discussion. However, the introduction of autonomous weapons raises novel issues of moral and political accountability.

The fact that each life has value, and the right to life is seen as the right to dignified life, presupposes that someone somewhere will internalise the cost of crossing the threshold of ending it; a human being or beings can and in many cases will assume or be assigned moral or political responsibility for its loss, even if it is otherwise legally permitted. In common parlance, it is on that person's conscience.

International law is not pacifist. War is sometimes a necessary evil. Human beings potentially have the capacity to appreciate the importance of the fact that even if war is necessary and legal, there is something wrong about it – there is a cost. This cost has traditionally been passed on through generations by those who know war, from parents to their children and grand-children; it is integrated in thousands of ways in our various cultures. We have a moral compass, which may lose its bearings at times, but which has the potential to corrected itself. Humans are the only species with this quality; neither animals nor machines have it.

On a very fundamental level, most people recoil from killing, and will go to great lengths to avoid it. "To prevent the scourge of war" was indeed one of the main objectives of the battle-weary leaders who gathered to form the United Nations.

It is difficult to see why the inner logic of machine learning over the decades would tend in the same direction, given that machines have no inherent ability to distinguish between good and evil, even if some such notions are initially programmed into some of the prototypes. The capacity to distinguish right and wrong is not part of their very nature.

Many discussions about autonomous weapons focus on the likelihood of compliance with legal or ethical standards in individual cases only, and not on their collective effect over time. If humanity now decides to allow fully autonomous weapons, it has the potential to stay with us for the centuries to come. To the extent that autonomous weapons are going to do the killing in our future wars, and humans increasingly defer to machines, there will be fewer incentives for individual human beings to work to prevent war as far as possible from occurring in the first place.

What is required is not only meaningful human control over every individual attack, and arguably that is not the main application of the term. The main concern animating the reaction against the idea of full machine autonomy is rather the prospect of losing meaningful human control over the long-term use of algorithms doing targeting in war in general. We may get so carried away by the short term and individualised advantages of machine targeting that we do not realise that we have lost control over the enterprise as a whole before it is too late. Preventing this can only be done by adopting a new legal norm.

Conclusion: what if fully autonomous weapons can save lives?

If machines, whether under meaningful human control or not, cannot do proper targeting, they should clearly not be used. For those machines under meaningful human control which can do proper targeting, the concerns raised above about the right to life and the right to dignity do not arise, and provided the rules of IHL are met, there can be no objection to their use. There may indeed be an obligation to use them, should they be available.

This leaves the case of machines where there is no meaningful human control, but they can ensure better targeting. Let us thus assume for a moment that the answer to the first question is 'yes' in respect to a fully autonomous weapon. Some kind of Turing test has been carried out,[2] and experts in targeting cannot, on a consistent basis, distinguish the accuracy and effects of machine and human targeting. If that point is reached, it will not take long before machines will surpass human targeting, so we can focus on that scenario.

To the extent that such a scenario is possible (which is contested), it is clear that a prohibition of full autonomy will potentially come at the expense of human lives that could otherwise have been saved – because a technology that offers lower risks is not being used.

This creates a potential dilemma for those who hold this position. Are they willing to argue that in such a case, the advantages of safer technology should be forfeited in the name of the protection of dignity?

It is indeed a central part of the principled opposition against fully autonomous weapons that the pursuit of social values, such as dignity, can result in us not using all the benefits that the technology otherwise offers. Many of our social practices are premised on the idea that quality of life is more important than taking all possible measures that may help to protect 'bare life'. For example, it remains a central tenet of the human rights project that torture is inherently wrong, and should be prohibited under all circumstances, even if it is possible to imagine a case where it can save lives. Likewise, even if mass surveillance can save lives in individual cases, for example by laying bare terrorist plots, accepting the practice should be viewed in the aggregate. There is a point beyond which the value we place on privacy will not allow it. The same reasoning, I would contend, applies to machines assuming the decision on life and death over humans.

2 See Wikipedia, Turing test: "The Turing test is a test (...) of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human", https://en.wikipedia.org/wiki/Turing_test.

Fully autonomous weapons do not entail a simple conflict between the right to life of civilians versus the dignity of those who may be targeted. As the above discussion of the "should they?" question demonstrated, the reality is more complex than that. Fully autonomous weapons present a potential threat to a range of rights, and to different right-holders. A number of conclusions reached above are worth repeating here.

Fully autonomous weapons present a threat to the right to life of those who are protected under IHL. Even if they pass the Turing test, there will still be errors regarding those who are protected, and there may be an accountability vacuum. And as we saw, a lack of accountability is a violation of the right to life. Moreover, it was argued that autonomous targeting is inherently arbitrary, because it does not entail human deliberation, and thus violates the right to life of those who are not protected and others who are affected. Autonomous targeting may also violate human dignity in a variety of ways, because it reduces human beings to objects that are to be destroyed.

Does the potential saving of lives justify these infringements? Human rights standards have little room for a utilitarian calculus based on abstract arguments about numbers of lives that may potentially be spared. Human rights are concerned with the fate of concrete individuals. The threats posed by the use of autonomous weapons to the rights to life and dignity are direct and concrete, while the consequences of not using it are indirect and at best speculative.

Even if it is possible that autonomous weapons may under certain circumstances reduce the risks of war and spare lives (and to what extent this is the case in practice is contested), there are consequently strong right to life and right to dignity arguments against the availability and use of such weapons if they have full autonomy.

Permitting fully autonomous targeting crosses a bridge of great importance. This may not be visible if only individual cases are considered – but the picture changes if the aggregate effect of changing the way in which targeting is conducted, and the exponential growth of computer capacity is borne in mind. The danger that we lose meaningful human control over targeting and thus war itself looms large.

Allowing the use of fully autonomous weapons in the name of saving unspecified lives somewhere in the future, not only in one particular case but as an acceptable practice, undermines the very reason why life is valuable in the first place.

Human-Machine Interaction in Terms of Various Degrees of Autonomy as well as Political and Legal Responsibility for Actions of Autonomous Systems

Neha Jain*

Introduction

The prospect of functioning autonomous systems in diverse areas of our lives poses unique challenges for the attribution of responsibility across a range of legal regimes: international humanitarian and criminal law regulating violations of the laws of war; tort law governing accidents caused by driverless cars; criminal law relating to harms resulting from employing robot carers and medical personnel. Of these, the laws relating to the deployment of autonomous weapons systems (AWS) have attracted the greatest scrutiny because of the sheer scale and extent of the potential harm and loss of human life.[1] The actions of an AWS, being partly of the character of a weapon and partly the character of a combatant,[2] will be enmeshed to a great extent within the actions of human agents acting together. In addition, by its very nature, the AWS will engage in conduct that is inherently unpredictable and dangerous. Should an AWS engage in conduct that violates the laws of war or commit an international crime due to malfunctioning, faulty programming, or incorrect deployment, who may be held responsible for this violation?

In the following sections, I will outline the primary challenges to the assignment of individual legal responsibility for the conduct of an AWS: the conceptual confusion surrounding the concept of machine autonomy; the range of actors involved in AWS conduct, leading to overlapping claims of responsibility; and the problem of effective control in light of the unpredictability of AWS conduct.

* Associate Professor of Law, University of Minnesota Law School. A more extensive version of these remarks can be found at N. Jain, Autonomous Weapons Systems: New Frameworks for Individual Responsibility, in N. Bhuta et al. (ed.), Autonomous Weapon Systems: Law, Ethics, Policy, Cambridge 2016, pp. 303-324.

1 See P. Scharre, Autonomous Weapons and Operational Risk, Report, Center for a New American Security, Washington 2016, https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk, p. 23.

2 H.Y. Liu, Categorization and Legality of Autonomous and Remote Weapons Systems, 94 International Review of the Red Cross 627 (2012) 629.

Conceptions of autonomy and human-machine interaction

The 'killer robot' is a popular cultural trope in science fiction novels and movies, media reports discussing AWS, and even campaign efforts to ban AWS.[3] This terminology plays on the sense of an AWS as a full-blown moral agent who is tireless, merciless, and capable of making its own decisions and executing them with precision. This strong sense of 'autonomy' finds some support in the technological literature. For example, Robert Sparrow defines an AWS as autonomous and thus morally responsible because it possesses the capacity to reason and to choose its own ends.[4] Similarly, for John Sullins an autonomous robot is a moral agent if it satisfies a number of criteria, including lack of control by another agent, effective accomplishment of tasks and goals, and intentional, deliberate behaviour.[5] However, neither Sparrow nor Sullins commit themselves to the proposition that such a fully autonomous system currently exists, or if it will ever exist.[6]

Indeed, given the current state of AWS technology, experts such as Noel Sharkey are highly critical of conceptualising the term 'autonomy' in robotics in the same way as used in philosophy, politics, individual freedom, or in common parlance.'[7] Sharkey describes an autonomous robot as one that "operates in open or unstructured environments".[8] Since the 'autonomous system' acts in a dynamic and unstructured environment based on feedback information from a number of sensors in its environment,[9] this inherently introduces uncertainty in its conduct.[10] Sharkey, following the US Navy's classification, categorises autonomous robots as (i) scripted, where the robots carry out a pre-planned script; (ii) supervised, in which robots perform planning, sensing and monitoring functions with the help of human operators, and (iii) intelligent, which are described rather ambiguously as those in which 'attributes of human intelligence' are used in software for decision making, problem solving, and perception and interpretation of information.[11]

3 See e.g. the campaign and various reports of Human Rights Watch named 'Killer Robots', https://www.hrw.org/topic/arms/killer-robots.
4 R. Sparrow, Killer Robots, 24 Journal of Applied Philosophy 62 (2007) 65-66.
5 J. P. Sullins, When is a Robot a Moral Agent?, 6 International Review of Information Ethics 23 (2006) 28-29.
6 Sparrow (n 4), p. 65-66; Sullins (n 5), p. 29.
7 N. Sharkey, Saying 'No!' To Lethal Autonomous Targeting, 9 Journal of Military Ethics 369 (2010) 376.
8 Ibid., pp. 376-377.
9 P. Asaro, On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making, 94 International Review of the Red Cross 687 (2012) 689.
10 International Committee of the Red Cross, Report of the ICRC Expert Meeting on Autonomous Weapon Systems: Technical, Military, Legal, and Humanitarian Aspects, 9 May 2014, https://www.icrc.org/eng/assets/files/2014/expert-meeting-autonomous-weapons-icrc-report-2014-05-09.pdf, p. 8.
11 Sharkey (n 7), p. 377.

The confusion engendered by equating 'autonomy' in machines with that in humans has led some scholars to abandon the term 'autonomy' altogether and instead use the term 'emergent' to describe AWS behaviour. Emergence refers to the ability of the system to engage in unpredictable, useful behaviour.[12] Emergent intelligence arises due to the interaction of thousands of parallel processes, where small program fragments examine and extract data from a range of inputs depending on the context.[13] Which inputs are considered relevant is contingent on pre-existing analytical structures and assumptions, which, in turn, are modified by new inputs as the system exhibits 'learning'.[14] Thus, the context used to extract and relate data may itself undergo changes under the influence of this new input. In this scenario, the basic statistics do not change, but the system will on occasion follow pathways that are unpredictable due to the constant slippage between the data and the context.[15]

This weaker sense of 'autonomy' or 'emergence' helps in narrowing down the focus of enquiry to the one respect in which AWS differ from conventional weapons systems: it is not because they possess moral agency similar to human agency, but because epistemic uncertainty is built into the very design of the system and is in fact crucial to its successful functioning. On what basis can a human agent be held responsible for such unintended and unanticipated conduct of an AWS, if it results in tremendous harm?

Control and causation

Legal responsibility across a range of domestic criminal law systems typically rests on the concepts of 'control' or 'causation', whereby the human agent who either causes or has control over the commission of the offence is said to be responsible for it. The concept of 'meaningful human control' has also been at the forefront of recent debates on responsibility for the conduct of an AWS,[16] having initially emerged during discussions on lethal autonomous weapons systems at the United Nations Convention on Certain Conventional

12 R. Calo, Robotics and the Lessons of Cyberlaw, 103 California Law Review 101 (2015) 119.
13 C.E.A. Carnow, Liability for Distributed Artificial Intelligences, 11 Berkeley Technology Law Journal 147 (1996) 159.
14 Ibid., p. 160.
15 Ibid., pp. 160–161.
16 See M.C. Horowitz and P. Scharre, Meaningful Human Control in Weapon Systems: A Primer. Working Paper, Center for a New American Security, 2015, https://s3.amazonaws.com/files.cnas.org/documents/Ethical_Autonomy_Working_Paper_031315.pdf.

Weapons in May 2014.[17] However, there is little clarity on how this phrase should be interpreted,[18] and various criteria have been proposed. For example, Article 36, the NGO that first introduced the concept, requires that in order to constitute meaningful human control, the attack must be executed by a human operator and that the operator and planners of the attack must have adequate contextual information over various aspects of the attack, and also be accountable for the outcome.[19] Similarly, the International Committee for Robot Arms Control proposes that the commander or operator should have "full contextual and situational awareness of the target area and be able to perceive and react to any change or unanticipated situations that may have arisen since planning the attack". In addition, there should be appropriate deliberation as to the nature of, and need for, the attack and its consequences, as well as the ability to call it off.[20] However, as experts such as Scharre have noted, such demanding criteria are unrealistic on the battlefield, and even some current weapons systems, the use of which is not controversial, would fail to meet them.[21]

The cognate concept of control in criminal law can help define more clearly both the kind of control that will be necessary for the attribution of responsibility in the event of harms resulting from AWS conduct, and also the challenges that will have to be overcome for its adoption. For instance, in civil law systems such as in Germany, in order to be responsible as the perpetrator of a crime, an individual should have 'control' over the act in question. To have control over the act means to hold in one's hands the elements constituting the offence (with the requisite intent).[22] This control can take different forms: direct domination over the act itself; control by the indirect perpetrator over the will of the direct perpetrator; functional domination of the participating joint actor in the case of co-perpetration.[23] The element of control signifies the ability of the perpetrator to execute or obstruct the commission of the offence according to his will.[24]

17 S. Knuckey, Governments Conclude First (Ever) Debate on Autonomous Weapons: What Happened and What's Next, Justsecurity.org, 16 May 2014, http://justsecurity.org/10518/autonomous-weapons-intergovernmental-meeting/.

18 Horowitz/Scharre (n 16), p. 6.

19 Article 36, Killer Robots: UK Government Policy on Fully Autonomous Weapons, April 2013, http://www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf.

20 F. Sauer, ICRAC statement on technical issues to the 2014 UN CCW Expert Meeting, 14 May 2014, http://icrac.net/2014/05/icrac-statement-on-technical-issues-to-the-un-ccw-expert-meeting/.

21 Horowitz/ Scharre (n 16), pp. 9, 11.

22 J. Wessels and W. Beulke, Strafrecht, Allgemeiner Teil: Die Straftat und ihr Aufbau, 42nd ed., Heidelberg 2012, p. 193.

23 Ibid.

24 Ibid.

Control is also central to another form of responsibility in domestic and international law: command responsibility. The doctrine of command or superior responsibility holds civilian and military superiors responsible for the unlawful conduct committed by their subordinates, and has three elements: (i) the existence of a superior-subordinate relationship where the superior has effective control over the subordinate; (ii) the requisite mental element, generally requiring that the superior knew or had reason to know (or should have known) of the subordinates' crimes, and (iii) the failure to control, prevent, or punish the commission of the offences.[25]

The superior-subordinate relationship includes both *de jure* and *de facto* command situations. The commander should have 'effective control' over the subordinate at the time of the commission of the act – that is, he should have the material ability to prevent and punish the offences.[26] Typically, a temporal coincidence between the control and the criminal conduct is required,[27] though some of the jurisprudence recognises instances of partial control if the superior fails to exercise his potential for full control in order to avoid personal responsibility.[28]

Identifying a single human agent who controls the actions of the AWS will prove challenging. Due to the unique features of an AWS, several human agents might be potential candidates for legal responsibility for its conduct. As UN Special Rapporteur Christopher Heyns notes, these could include "the software programmers, those who build or sell hardware, military commanders, subordinates who deploy these systems and political leaders [who authorize them]".[29] On the face of it, the programmer or software developer is too remote from the scene of the crime to control the AWS's conduct at the time of the commission of the offence. This is especially true given the fact that the AWS is designed to apply non-deterministic decision-making in an open and unstructured environment. Thus, not only will pre-set rules designed by the programmer not be capable of capturing the complete range

25 A.M. Danner and J.S. Martinez, Guilty Associations: Joint Criminal Enterprise, Command Responsibility, and the Development of International Criminal Law, 93 California Law Review 75 (2005) 122.

26 Ibid., p. 130.

27 International Criminal Court, Prosecutor v Jean-Pierre Bemba Gombo (Bemba), Case no. ICC-01/05-01/08, Decision Pursuant to Article 61(7)(a) and (b) of the Rome Statute on the Charges of the Prosecutor against Jean-Pierre Bemba Gombo (Pre-Trial Chamber II, 15 June 2009), para. 418.

28 I. Bantekas, The Contemporary Law of Superior Responsibility, 93 American Journal of International Law 573 (1999) 580 (citing the *Yamashita* and High Command cases).

29 C. Heyns, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, UN Doc. A/HRC/23/47, 2013, para 77.

of scenarios that the AWS will encounter, but the AWS will also be able to adapt the means and methods it uses to achieve programmer-designed ends. This will inevitably lead to uncertainty and error, which cannot be fully predicted or controlled by the programmer.[30]

Similarly, the field commander who deploys the AWS may cause or control the AWS's conduct only to the extent that he may be in a position to terminate the immediate action and prevent future actions. The crucial feature of AWS functioning is its ability to individualise target selection and the means of carrying out an attack on the basis of changing circumstances.[31] Thus, the field commander may not be well equipped to direct, predict, or second-guess the decision-making operation carried out by the AWS in real time. This would also hold true of the civilian or military superior who is doubly distanced from the real-time operation of the AWS.

One possible solution to this problem is to shift the focus of when control is relevant. Thus, instead of evaluating the locus of control at the time that the AWS is deployed, the law should concentrate on the point when the process to develop the AWS to act within the bounds of specific program features is undertaken, and the decision and to deploy it in certain field operations is made.[32] This shift departs from the jurisprudential requirement of temporal coincidence for attribution of command responsibility. However, the rationale behind the temporal requirement is that the field or operational commander is expected to exercise both *de facto* and *de jure* control over the human agent who he trains and deems fit for deployment and over whom he exercises disciplinary authority. In the case of the AWS, the analogous *de facto* and *de jure* control will be exercised by the commander who reviews the ability of the AWS to be able to perform the tasks assigned to it within the limits of the law and gives the authorisation to deploy it for a certain operation. The commander may still be considered to 'control' its actions due to his role in setting (or failing to set) the conditions under which the AWS operates.

30 On the unpredictable functioning of the AWS see O. Gross, When Machines Kill: Criminal Responsibility for International Crimes Committed by Lethal Autonomous Robots, unpublished manuscript, 9 October 2014 (on file with the author).

31 M. Wagner, The Dehumanization of International Humanitarian Law: Legal, Ethical, and Political Implications of Autonomous Weapon Systems, draft paper, www.law.upenn.edu/live/files/4003-20141120---wagner-markus-dehumanizationpdf., p. 13.

32 See e.g. the suggestion by M. Schulzke, Autonomous Weapons and Distributed Responsibility, 26 Philosophy & Technology 203 (2013) 213-215.

This change in focus will only prove workable if the commander (and the operator) have adequate information about the nature, design, and functioning of the weapon to be able to take an informed decision regarding whether it is capable of being deployed in a lawful manner. Additionally, rigorous testing and inspection of the weapon and training of operators will be essential in order to ensure that they have the requisite information and capability to exercise control, and abort the mission if necessary.[33] Indeed, as some scholars have argued, decision-making in the development and procurement phase might be the most crucial place for considering compliance with the laws of war. This, in turn, will have implications for the locus of control.[34] However, several additional factors may limit this control, including the commander's ability to actually prevent or punish the conduct, and the epistemic uncertainty associated with AWS conduct.[35]

Foreseeability and risk

Notwithstanding proof of 'control', attribution of legal responsibility has yet another element that is difficult to satisfy in the face of unpredictability: the mental state accompanying the conduct that is typically required to fulfil the preconditions for both civil and criminal responsibility.

Under current international criminal law as provided under the Rome Statute of the International Criminal Court, criminal responsibility is predicated on the perpetrator's intent or knowledge with respect to the elements of the respective offence. Under Article 30 of the Rome Statute, in order to act with 'intent', the perpetrator must mean to engage in the conduct and mean to cause or be aware that the consequence will occur in the ordinary course of events. 'Knowledge' requires awareness that a circumstance exists or a consequence will occur in the ordinary course of events. Given the inherently unpredictable nature of AWS conduct, the harm caused by its actions is unlikely to fit either of these mental standards. Command responsibility lowers this mental element considerably. Thus, under Article 28 of the Rome Statute, a civilian superior must have either known or "consciously disregarded

33 See Horowitz/Scharre (n 16), pp. 14-15, proposing similar criteria for 'meaningful human control'.

34 G.S. Corn, Autonomous Weapon Systems: Managing the Inevitability of 'Taking the Man Out of the Loop, in Nehal Bhuta et al. (ed.), Autonomous Weapon Systems: Law, Ethics, Policy, Cambridge 2016, pp. 209-242.

35 H.Y. Liu, Refining Responsibility: Defining Two Types of Responsibility Issues Raised by Autonomous Weapons Systems, in Nehal Bhuta et al. (ed.), Autonomous Weapon Systems: Law, Ethics, Policy, Cambridge 2016, pp. 325-344.

information which clearly indicated, that the subordinates were committing or about to commit such crimes". The parallel standard of "owing to the circumstances, should have known" for military commanders is still lower and amounts to negligence.

In the context of responsibility for the conduct of AWS, two questions arise from these reduced standards: first, even if we accept the lower level of recklessness or negligence, will this enable us to assign responsibility for inherently unpredictable AWS conduct to the commanders/civilian superiors? And if yes, should we lower the mental element more broadly to capture the conduct of other actors such as deploying soldiers or field officers?

The concepts of 'recklessness' and 'negligence' are not interpreted uniformly across legal systems and in international criminal law. According to the influential definition proposed in section 2.02(2) of the United States Model Penal Code (MPC), for instance, recklessness and negligence both involve the individual disregarding a substantial and unjustifiable risk with respect to the material elements of an offence. However, while a reckless defendant is aware of the risk, the negligent defendant is not, but should have been.[36] What exactly the defendant must be aware of, or believe in relation to the risk, is highly disputed. Should the defendant only be aware of the risk, with the question of whether it is substantial and unjustifiable being determined by the adjudicator? Alternatively, should he be subjectively aware that the risk is both substantial and unjustifiable? There is no consensus on these issues either within the academic community[37] or in positive law.[38]

Relatedly, how concrete or exacting the defendant's conception of the risk needs to be is also unclear. Should the defendant foresee the exact harm that occurs, roughly the same type or category of harm,[39] or simply be aware that there is a substantial and unjustifiable risk of a 'dangerous' occurrence?

36 D. Husak, Negligence, Belief, Blame and Criminal Liability: The Special Case of Forgetting, 5 Criminal Law & Philosophy 199 (2011) 200.

37 See K.W. Simons, When Is Negligent Advertence Culpable, 5 Criminal Law & Philosophy 97 (2011) 112; D.M. Treiman, Recklessness and the Model Penal Code, 9 American Journal of Criminal Law 281 (1981) 365; L. Alexander, Insufficient Concern: A Unified Conception of Criminal Culpability, 88 California Law Review 931 (2000) 934-935.

38 Husak (n 36), p. 208.

39 Moore and Hurd argue that the defendant should be aware that "there is some risk of roughly the type that was realized"; see M.S. Moore and H.M. Hurd, Punishing the Awkward, the Stupid, the Weak, and the Selfish: The Culpability of Negligence, 5 Criminal Law & Philosophy 147 (2011) 156.

If we adopt a broader notion of recklessness, where the defendant is responsible even when he is not conscious of the exact risk his conduct creates, this could capture liability for the harm caused by an AWS. Thus, even if the actions of the AWS are unpredictable and even if the defendant was unaware of the exact nature of the risk of harm posed by the conduct, he could still be deemed reckless and, hence, liable for the harm. However, liability for recklessness will still require showing some level of subjective awareness of some kind and degree of risk, and depending on the extent of epistemic uncertainty associated with AWS conduct, this might be difficult to prove in individual cases.

Negligence liability for AWS conduct would alleviate this problem to some extent. Negligence as a culpable mental state expands the responsibility regime, since the defendant's inadvertent risk creation suffices for liability. The negligence standard under the MPC is higher than ordinary or simple negligence in tort. Thus, the defendant must deviate from the standard of care required of the reasonable person in a manner that is extreme or gross enough to warrant criminal liability.[40] In the context of harmful AWS conduct, the adjudicator will have to determine whether the commander, field officer, or deploying soldier, respectively, should have been aware of a substantial and unjustifiable risk of harm resulting from AWS conduct, and if, given the circumstances and their individual knowledge, their failure to advert to this risk constituted a gross deviation from the standard of care expected of a reasonable person in their situation.

Depending on the extent to which the defendant's 'situation' is taken into consideration in the case of the commander, field officer, or deploying soldier, there could be a potential exoneration from liability if the AWS behaves in an entirely unprecedented fashion that no reasonable individual who did not have some knowledge of the technical details of its operation could have foreseen. This is more likely to be the case the further one moves down the chain of command, especially to the deploying soldier who is unlikely to be aware of the exact mode of functioning of the AWS, the nature of the safety precautions that have been built into the system, and when the machine may deviate from its standard operating procedure and behave in unexpected ways that may cause unforeseen harm.

As the earlier analysis should have made clear, lowering the mental requirements to recklessness and/or negligence will not capture all of the instances of AWS functioning gone awry, but it will certainly broaden the potential scope of liability to cover individual cases. Nevertheless, there are important policy questions that will have to be addressed as to the

40 Husak (n 36), p. 202.

costs of this proposal: most crucially, any steps to embrace the negligence standard more widely must be attentive to the scepticism towards negligence liability in most domestic criminal legal systems.[41]

The above analysis mostly relates to the criminal responsibility of civilian and military superiors, field officers, and the individuals who are in charge of deploying the AWS. Other actors, such as software developers, programmers and manufacturers, will generally be too remote from the actual deployment and from the scene of the crime in order to establish control for the purposes of criminal responsibility. They could nonetheless be held liable under civil law principles of product liability based on a standard of negligence and/or strict liability for harm that results from malfunctions, or from poor or error-prone software or hardware design features of the AWS.[42]

The foreseeability issue will, however, continue to be relevant even for civil/tortious liability.[43] In legal systems such as the in United States, and in most common law jurisdictions, negligence liability has traditionally been based on the concept of reasonable foreseeability. Thus, if the injury that is caused by the defendant was not reasonably foreseeable, he will not be responsible for the injury.[44] The centrality of foreseeability in negligence liability has been emphasised for evaluating both whether the defendant breached his duty of care, and to establish proximate cause.[45] Going even further, 'strict liability' cases of tort liability also retain the requirement of foreseeability, and courts often require proof of the foreseeability of one or more of three elements: the kind or type of risk of harm, the person likely to be harmed, and the manner of harm.[46]

41 Ibid., p. 203; see also the discussion and references at Moore/Hurd (n 39), p. 150.

42 Schulzke (n 32), p. 214.

43 See Calo (n 12), p. 141.

44 See the discussion at B.C. Zipursky, Foreseeability in Breach, Duty, and Proximate Cause, 44 Wake Forest Law Review 1247 (200) 1256; as he notes, the Restatement (Third) of Torts proposes a modified standard of foreseeability than that endorsed in the classical doctrine.

45 W.J. Cardi, Reconstructing Foreseeability, 46 Boston College Law Review 921 (2005) 925–927.

46 C.E.A. Carnow, The Application of Traditional Tort Theory to Embodied Machine Intelligence, The Robotics and the Law Conference, Center for Internet and Society, Stanford, 13 April 2013, citing D.A. Fischer, Products Liability-Proximate Cause, Intervening Cause, and Duty, 52 Modern Law Review 547 (1987) 553; see also Calo (n 12), p. 141.

The importance placed on foreseeability for tortious liability[47] will pose difficulties in holding the programmer, manufacturer, or developer liable for the same reasons that were relevant to criminal responsibility. An AWS is designed to observe, decide, and act in a manner that responds to the pressures of evolving and complex environments. Rather than a design defect, the ability to function in an unpredictable fashion is built into the nature of the system and gives it value.

Conclusion

Legal responsibility for unlawful AWS conduct requires academics, practitioners, and policy makers to grapple with the concept of 'autonomy' that is salient for the attribution of liability. Given the current state of technology, autonomy in AWS should not be confused with moral agency in human beings. Rather, autonomy in the machine context relates to the ability of the system to rely on non-deterministic reasoning in order to operate in unstructured and dynamic environments.

This necessarily entails a level of epistemic uncertainty and unpredictability associated with AWS conduct. For civil as well as criminal responsibility, the concepts of 'control' and 'foreseeability' point to issues that will be central to developing a paradigm for legal responsibility for AWS conduct, and how the attendant risks should be allocated amongst the different actors developing, supervising, and operationalising the use of AWS.

47 On the importance of foreseeability for tort liability see D. Owens, Figuring Foreseeability, 44 Wake Forest Law Review 1277 (2009) 1281-1290.

The Distraction of Full Autonomy and the Need to Refocus the CCW Laws Discussion on Critical Functions

Chris Jenks*

Introduction

> "*During the whole debate on technical issues of lethal autonomous weapons systems, the notion of autonomy and its definition was at the center of interest.* It became quite obvious that there is no ready-made, generally accepted definition of what is an 'autonomous system' and as to where to draw the line between 'autonomous' and 'automatic' or 'automated'." [1]

That apt description summarises the results of the first meeting of experts on lethal autonomous weapons systems (LAWS), held in Geneva, Switzerland, in May 2014, under the auspices of the United Nations' Convention on Certain Conventional Weapons (CCW). The international community held the third such meeting in May 2016, at which States Parties agreed to recommend to the upcoming CCW Fifth Review Conference that a group of governmental experts (GGE) meet in 2017-2018. With the amount of time and effort already expended, and likely with (at least) two more years of meetings to come, it seems appropriate to take stock of the LAWS discussions to date.

Bluntly stated, the LAWS discussions have been confused, not constructive, and largely for the same definitional reasons identified two years ago. This section attempts to address the question why the dialogue has proceeded as it has, and proposes how it should proceed at the GGE meetings.

Attention and concern on LAWS has rapidly gained momentum, leading to CCW States Parties agreeing to convene the informal discussions of the last three years. Yet, despite those meetings, there still is not even consensus about what is being discussed, due, at least in part,

* Assistant Professor of Law, SMU Dedman School of Law, Dallas, Texas. This article supplements a presentation delivered as part of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons as part of the Convention on Certain Conventional Weapons treaty.

1 M. Biontino, Summary of Technical Issues, CCW Expert Meeting Lethal Autonomous Weapons Systems, 13-16 May 2014, http://www.genf.diplo.de/Vertretung/genf/en/02/statements-en.html. (emphasis added).

to the varied understandings and meanings of autonomy. This section focuses on the problems created by framing the LAWS discussion in terms of full autonomy and suggests that CCW States Parties refocus on the critical functions of selecting and engaging targets.

The results of the LAWS discussions to date are akin to a car racing into a cul-de-sac. That the car has reached the cul-de-sac quickly is meaningless. And once there, the options are limited and unproductive – stop, withdrawal, or drive in circles. In order for that to change, the LAWS discussion must move beyond and away from the conceptual framework of full autonomy. That framework both confuses and distracts – and has driven the dialogue in an unproductive circular direction. Focusing on the critical functions of selecting and engaging of targets may facilitate progress, and is consistent with CCW's purpose "to ban or restrict the use of specific types of weapons that are considered to cause unnecessary or unjustifiable suffering to combatants or to affect civilians indiscriminately".[2]

To explain why framing the discussion in terms of full autonomy dooms the LAWS discussion to be perpetually circular in nature, Part I of this paper explores the challenges in trying to define and otherwise categorise system autonomy. This includes why distinguishing systems based on whether a human is in, on, or out of the loop, or whether a system is automated, automatic, or autonomous, is only of minimal, descriptive, utility. Part II then applies the concept of autonomy to weapons systems, focusing first on the reality that weapons systems capable of selecting and engaging targets without further human intervention have existed for decades, and then on explaining why full autonomy is a distractor. Part III then emphasises the need to refocus the LAWS conversation within the purpose of the CCW.

Autonomy

As noted at the outset, a threshold challenge in discussing LAWS is that there are wildly varied understandings of what is meant by autonomy in general, let alone as applied to weapons systems. Given autonomy's complex nature, this should not be surprising. But these different understandings set the conditions for a dialogue bordering on incoherence. So much so that it would be tremendous progress for the international community if there

2 See Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to be Excessively Injurious or to Have Indiscriminate Effects, 10 October 1980, 1342 U.N.T.S. 137, 19 I.L.M. 1523.

was a complete and utter *lack* of consensus regarding whether to develop and employ LAWS, but agreement as to what was meant by LAWS.[3] But as of now, we cannot even agree on what we are discussing.

The United Nations Institute for Disarmament Research succinctly explained how the discussion of LAWS

> presently lacks focus, tacking between things (for example, drones, robots and systems), a characteristic (autonomy) and uses (defensive measures? Targeting? Kill Decisions?) in an inconsistent and often confusing way. One of the reasons there are so many different terms being proposed as the object of discussion is that some actors are trying to capture a mix of variables of concern (such as lethality or degrees of human control), while others are talking about more general categories of objects.[4]

A constructive LAWS dialogue requires a shared and coherent understanding of machine or system autonomy. Thus far, the international community has neither, and perhaps even worse, continues to engage in overly broad and conceptually confusing inquiries. In order to recognise the inherent futility of these inquiries, the discussion could draw from the U.S. militaries' failed efforts in conceptualising and explaining autonomy.

3 See for example the representative of Human Rights Watch at the 2016 CCW LAWS Experts Meeting: "We are talking about future weapons systems that once initiated, using sensors and artificial intelligence, will be able to operate without meaningful human control, that will be able to select and engage targets on their own, rather than a human making targeting and kill decisions for each individual attack." Statement by Stephen Goose of Human Rights Watch, General Exchange of Views, Informal Meeting of Experts on Lethal Autonomous Weapons Systems Convention on Conventional Weapons, Geneva, 12 April 2016, http://www.unog.ch/80256EDD006B8954/(httpAssets)/252007F8C3EB3E1EC1257FAE002F4DE5/$file/HRW+intervention+Goose+12+April+2016.pdf (emphasis added). The International Committee of the Red Cross, on the other hand, reiterated its view that "[s]ome weapon systems in use today can select and attack targets without human intervention". Statement of the International Committee of the Red Cross, Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), 11 April 2016, http://www.unog.ch/80256EDD006B8954/(httpAssets)/9324B81015529E3DC1257F930057AF12/$file/2016_LAWS+MX_GeneralExchange_Statements_ICRC.pdf (emphasis added). Thus, under HRW's view, the CCW LAWS discussion is about future, whereas according to ICRC, at least some current weapons systems would, by definition, be considered as LAWS and thus factor into that same discussion.

4 United Nations Institute for Disarmament Research, *Framing Discussions on the Weaponization of Increasingly Autonomous Technologies*, 3 (2014)

In 2012, the U.S. Defense Science Board issued a report, "The Role of Autonomy in [Department of Defense] Systems".[5] As the report explains, what had been occurring was that different U.S. military services (Army, Navy, and Air Force) were "making significant investments of time and money to develop definitions of autonomy", yet "[t]he competing definitions for autonomy have led to confusion". And that confusion "may be contributing to fears of unbounded autonomy".[6] The end result was "a waste of both time and money spent debating and reconciling different terms"[7] which were "irrelevant to the real problems".[8]

While the desire to define autonomy is natural and seems reasonable, such efforts will inevitably be counterproductive.[9] The U.S. militaries' definitional efforts "unsatisfactorily (...) tried to express autonomy as a widget or discrete component (...)".[10] Attempts to develop 'autonomy roadmaps' based on trying to correlate certain levels and types of computer functions needed for a certain level of autonomy were equal parts well-intended and ill-advised. The concept of autonomy is neither usefully nor helpfully thought of in levels. This is because of, among other reasons, the dynamic nature of functions within a system. Many functions can be executed "concurrently as well as sequentially (...) and can have a different allocation scheme to the human or computer at a given time".[11]

Autonomy is better thought of across not one but several spectrums.[12] And within each spectrum, the amount or quality of machine autonomy varies and changes as the system operates. As a result, plotting autonomy as a linear and single axis progressively and discretely demarcated by whether humans are in, on, or out of a functional loop both oversimplifies and misrepresents. It amounts to a form of conceptualising autonomy in levels, which, as discussed, is neither useful nor helpful beyond serving as a descriptor. Nor is there consensus on where and how to delineate in, on, and out of the loop.

5 U.S. Department of Defense, Defense Science Board, Task Force Report: The Role of Autonomy in DoD Systems, July 2012, http://fas.org/irp/agency/dod/dsb/autonomy.pdf, p. 1.
6 Ibid; the report acknowledged that "the word 'autonomy' often conjures images in the press and the minds of some military leaders of computers making independent decisions and taking uncontrolled action".
7 J.M. Bradshaw et al., The Seven Deadly Myths of "Autonomous Systems", Human-Centred Computing, May/June 2013, p. 57, www.jeffreymbradshaw.net/publications/IS-28-03-HCC_1.pdf.
8 Ibid.
9 Ibid.
10 U.S. Department of Defense (n 4).
11 U.S. Department of Defense (n 4).
12 Three such spectrums are the nature of the human-machine command and control relationships, the complexity of the machine and the type of decision being automated. See P. Scharre, Between a Roomba and a Terminator: What is Autonomy, War on the Rocks, 18 February 2015, http://warontherocks.com/2015/02/between-a-roomba-and-a-terminator-what-is-autonomy/.

Similarly, attempting to broadly differentiate machine functions as being either automatic, automated, or autonomous lacks practical utility. These terms may be used to understand as one of the spectrums through which we conceptualise autonomy, the complexity of the machine. This spectrum ranges from the lower end, automatic, to automated, to autonomous at the higher end.[13] But again, this is only one spectrum. And the utility of that one spectrum is limited, as there are no clear boundaries between automatic, automated, or autonomous,[14] a challenge which CCW States Parties have already encountered.[15]

Consider the following assessment of a household cleaning device, the Roomba robotic vacuum cleaner:

> "The Roomba must navigate a house full of obstacles while ensuring that the carpet is cleaned. (...) The Roomba user provides high-level goals (vacuum the floor, but don't vacuum here, vacuum at this time of day, etc.). The Roomba must make some choices itself (how to identify the room geometry, avoid obstacles, when to recharge its battery, etc.). The Roomba also has some automated behaviour and encounters situations it cannot resolve on its own (e.g., it gets stuck, it can't clean its own brushes, etc.). Overall, the Roomba has marginal autonomy, and there are numerous situations it cannot deal with by itself. It is certainly not intelligent. However, it does have basic on-board diagnostic capability ("clean my brushes!") and a strategy for vacuuming a room about whose size and layout it was initially ignorant."[16]

Where should the Roomba be placed within the human in, on, or out of the loop? Is the Roomba automated? Autonomous?[17] Similarly, where within the loop should a household thermostat or microwave oven be placed? The answer, which further illustrates that we cannot draw system-wide conclusions from this spectrum, is that we do not know

13 Ibid.

14 Ibid.

15 See above; additionally, the focus on autonomously performed functions works toward the effect that such a system would be the subject of the CCW discussion, while one which automatically selected and engaged targets, or was automated, would not.

16 C.R. Frost, Challenges and Opportunities for Autonomous Systems in Space, in: National Academy of Engineering (ed.), Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2010 Symposium, 2011, pp. 89-102.

17 Similarly consider driving a car, its features and functions, and which are activated by the human driver vs. by the car itself: "Most cars today include anti-lock brakes, traction and stability control, power steering, emergency seat belt retractors and air bags. Higher-end cars may include intelligent cruise control, automatic lane keeping, collision avoidance and automatic parking." P. Scharre and M. Horowitz, An Introduction to Autonomy in Weapon Systems, Center for a New American Security, February 2015, https://www.cnas.org/publications/reports/an-introduction-to-autonomy-in-weapon-systems.

without having more information about the system. Certain thermostats and microwave ovens would likely be considered automated, others capable of sensing and adjusting their operation may be autonomous.[18] But again the lines are fuzzy, and machine complexity is but one spectrum.

Ultimately,

> "autonomy isn't a discrete property of a work system, nor is it a particular kind of technology; it's an idealized characterization of observed or anticipated interactions between the machine, the work to be accomplished, and the situation. To the degree that autonomy is actually realized in practice, it's through the combination of these interactions".[19]

Autonomy is better thought of as "a capability of the larger system enabled by the integration of human and machine abilities".[20] This approach recognises that the operation of all machines requires some degree of human involvement. That means that there are no fully autonomous systems.[21] The point of autonomy being bounded extends to weapons systems as well – there is no such thing as a fully autonomous weapon.[22]

18 Asaro distinguishes automated from autonomous on the grounds that unsupervised automated systems "involve repetitive, structured, routine operations without much feedback information (such as a dishwasher)", while autonomous systems operate in "dynamic, unstructured, open environments based on feedback information from a variety of sensors (such as a self-driving car)"; P. Asaro, On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision Making, 94 International Review of the Red Cross 687 (2012) 690. Assuming one can apply the approach to weapons systems, it is unclear to which extent LAWS are more comparable to the dishwasher than the self-driving car, let alone why the focus is on overall categorisation and not critical functions such as engagement.

19 Bradshaw (n 6); another way of thinking about autonomous machines in is terms of the extent of self-directedness and self-sufficiency.

20 U.S. Department of Defense (n 4).

21 Ibid.

22 As the President of the ICRC observed in 2011 (referring to 'truly' as opposed to 'fully' autonomous weapons), "such systems have not yet been weaponised. Their development represents a monumental programming challenge that may well prove impossible." J. Kellenberger, International Humanitarian Law and New Weapon Technologies, statement at the 34th Round Table on Current Issues of International Humanitarian Law, 8 September 2011, https://www.icrc.org/eng/resources/documents/statement/new-weapon-technologies-statement-2011-09-08.htm.

Indeed, as one commentator observed, "the question of when we will get to 'full autonomy' is meaningless. There is not a single spectrum along which autonomy moves. (...) A better framework would be to ask what tasks are done by a person and which by a machine".[23] Utilising that framework would in turn better focus the inquiry on the critical functions of selecting and engaging targets.

Lethal Autonomous Weapons Systems

The LAWS conversation must acknowledge that weapons systems capable of selecting and engaging targets without further human intervention have existed for decades. Framing the issue in terms of fully autonomous systems is not such an acknowledgement, it is rather an attempt to avoid that underlying reality.[24] The inescapable problem is that the concern for LAWS is grounded in systems capable of determining what to shoot or fire at and then shooting or firing. And such systems already exist.[25] Unless and until the international community can identify what it is beyond autonomy in the critical function of selection and engagement of targets that is troubling, then there is little advantage in framing the discussion in terms of full autonomy.

Examples of current LAWS

In a report following a 2014 experts meeting, the International Committee of the Red Cross (ICRC) provided a sampling of existing weapons systems with autonomy in the critical functions of acquiring, tracking, selecting, and attacking targets:[26]

23 See Scharre (n 12).

24 Framing the LAWS discussion in terms of future, full autonomy allows both States Parties which currently field what would otherwise qualify as LAWS, and civil society groups advocating States Parties to support a ban, to not talk about current weapons systems. The problem is that if the LAWS discussion is to focus on weapons systems capable of selecting and engaging targets without further human intervention, then current systems will have to be included.

25 See Scharre/Horowitz (n 17), at Appendix b, (detailing that over thirty countries have employed fifteen different autonomous weapons systems dating as far back as 1980).

26 International Committee of the Red Cross, Expert Meeting: Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects, March 2014, https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014. The vast majority of these weapons systems are designed to target material, aircraft, vessels at sea, and inbound missiles. Other estimates are that "[a] s many as 40 nations are currently developing military robotics", and "some weapons already in use may be considered autonomous"; S. Gross, The U.S. Should Oppose the U.N.'s Attempt to Ban Autonomous Weapons, The Heritage Foundation, 5 March 2015.

- Patriot surface-to-air missile system; a missile defence system that automatically detects and tracks targets before firing interceptor missiles;[27]
- Aegis Weapon System; a ship-based system combining radar to automatically detect and track targets with various missile and gun systems;[28]
- Phalanx Close-in Weapon System; a ship-based 20 mm gun system that autonomously detects, tracks, and attacks targets;[29]
- The Goalkeeper Close-in Weapon System; "an autonomous and completely automatic weapon system for short-range defence of ships against highly manoeuvrable missiles, aircraft and fast manoeuvring surface vessels";[30]
- Counter Rocket, Artillery, and Mortar System; a land-based fixed weapons system that employs the same technology as the Phalanx Close-in Weapon System to target and attack rockets, artillery, and mortars;[31]
- Iron Dome; a ground-based air defence system which automatically selects targets and fires interceptor missiles;[32]
- NBC MANTIS (Modular, Automatic and Network-capable Targeting and Interception System); an automated ground-based air defence system using 35 mm guns to automatically target rocket, artillery, and mortars.[33]

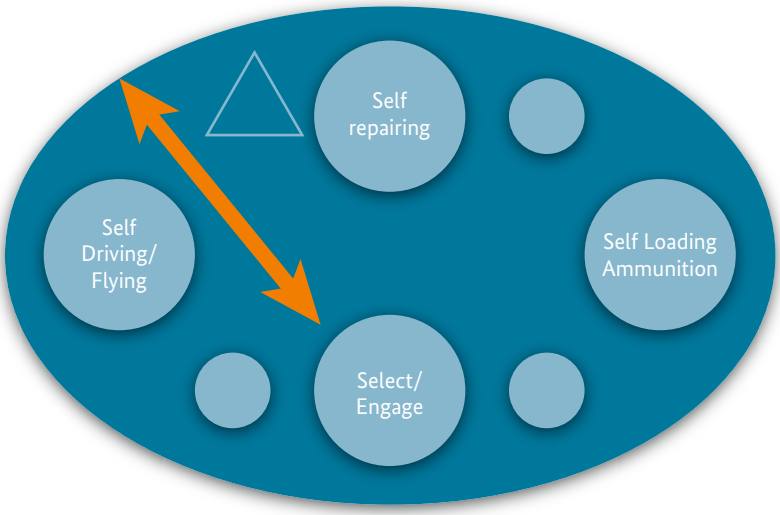
The majority of these weapons systems are not new, to varying degrees they have been in use for decades. As a result, it is challenging to assemble a coherent contemporary argument as to why 1980s weapons systems employed with minimal issues are now problematic. Perhaps focusing on the looming future and the prospect of fully autonomous weapons is, superficially anyway, easier than trying to articulate a retrospective argument. But framing the LAWS discussion in terms of full autonomy needlessly distracts and confuses.

27 Ibid., p. 67; introduced in the late 1970s, the Patriot is employed by at least sixteen countries.
28 Ibid; introduced in 1978, Aegis is employed by at least five countries.
29 Ibid; introduced in 1980, Phalanx is employed by at least twenty-five nations; see Raytheon, Phalanx Close in Weapon System, http://www.raytheon.com/capabilities/products/phalanx/.
30 Ibid., p. 65; Goalkeeper is operational in at least eight navies; see Thales, Goalkeeper Close in Weapon System, https://www.thalesgroup.com/en/goalkeeper-close-weapon-system.
31 Ibid.
32 Ibid.
33 Ibid.

Distraction of fully autonomous weapons

What is it about autonomy, above and beyond machines selecting and engaging targets without human intervention, that is worrisome, and is CCW the appropriate venue for those concerns? Recognising that full autonomy, if it is possible, would be unbounded, and depicting that, in two dimensions, it is challenging, consider full autonomy as the oval below, comprised of any number of functions (presumably infinite), depicted by the other ovals.

Full Autonomy



There is space, an autonomy 'delta'/Δ, between full autonomy and autonomy in the function of selecting and engaging targets. Among other functions, this delta could be manifested by, for example, logistics trucks[34] or aircraft[35] capable of self-driving or flying, of recharging or reloading themselves, or performing internal diagnostic assessments and repairs.

34 See L.M. Bacon, Unmanned Vehicles Heading for Battlefield, Army Times, 2 September 2013 (describing military plans for driverless trucks).

35 See J. Gould, Army Seeks to Cut Casualties with Unmanned Delivery Craft, Army Times, 8 December 2014; A. Mehta, Sikorsky Plans First Flight of Autonomous Black Hawk, Army Times, 14 May 2014 (describing military plans for unmanned aerial delivery systems, which would not be remotely piloted but autonomously perform flight functions).

To ground the LAWS debate in full autonomy requires articulating what is concerning about systems *beyond* the capability to select and engage targets without human intervention.[36] Otherwise, as this section advocates, the discussion should focus on the critical functions of selecting and engaging targets.

Consider a hypothetical weapons system "A", which performs all functions autonomously *except* selection and engagement of targets.

Now consider the inverse of A, hypothetical system "B", in which humans perform virtually all the functions, but B, not a human operator, selects and engages targets.

Which of the functions the different weapons systems perform are concerning, and by what criteria? None of the functions A performs without human intervention potentially causes unnecessary or unjustifiable suffering to combatants, or affect civilians indiscriminately. And those functions the system performs are not a 'specific weapon'. These non-weapons functions are a reminder that autonomy is merely a technology descriptor.

The point of introducing systems A and B is to underscore that while there may well be concerns about autonomous non-weapons functions (the self-driving, the self-loading, etc.), they are not most appropriately addressed within the CCW, a weapons treaty regime. Fully autonomous systems would constitute systems A+B. But what is it beyond autonomously performed weapons functions that is potentially concerning? Stated otherwise, why not limit the LAWS discussion to those critical weapons system functions? Doing so makes the conversation much more manageable and in line with the CCW.[37] As has been

36 Otherwise, the argument could – and this section contends that it should – focus on weapons systems capable of selecting and engaging targets without further human intervention.

37 This paper does not mean to suggest that refocusing the CCW LAWS discussion on critical functions would yield clarity or rapid resolution, far from it. Indeed, there are a number of very challenging questions to be sorted within the context of critical functions – what it means to select, what it means to engage, and what and how we think of human control – that are quite daunting. Yet, consider that three years into the CCW LAWS discussions there has been no mention of what is meant by select or engage. But framing in terms of critical functions is much closer to, if not a bounded inquiry, and States Parties are more likely to at least agree on what it is that is being discussed. That by itself would constitute a sizable step forward from the status quo.

demonstrated, a dialogue where LAWS are framed in full autonomy tends to drift between science, including artificial intelligence[38] and the concept of singularity,[39] and science fiction, including robots somehow countermanding human-imposed operating constraints.[40]

We do not know what full autonomy means or if it is even achievable, and if so, when. This is a challenging foundation upon which to engage in constructive dialogue. Moreover, why does autonomy other than in selection and engagement of targets matter, at least within the CCW construct?

CCW

International consensus is of course challenging to obtain in any area, but by framing LAWS in terms of full autonomy, the discussion undermines itself. Because of the inability to reconcile potential, future, fully autonomous systems with long-standing weapons systems capable of selecting and engaging targets without further human intervention, the dialogue thus far has been limited to general terms. But this lack of specificity has either set the conditions for, or allowed the discussion to become circular, depending on one's individual perspective.

This is reflected in what has transpired at the CCW since the treaty body took on LAWS in 2013. In 2014, the majority of the debate was spent discussing how to define autonomy.[41] Not only was an agreement on the definition not reached, a majority of the delegations

38 McCarthy, the professor who developed the term artificial intelligence [AI] in 1956, defined it as "science and engineering of making intelligent machines, especially intelligent computer programs"; see N. Bostrom, Superintelligence Paths, Dangers Strategies, 2014. But AI means different things to different researchers.

39 While it is unclear whether full autonomy could ever exist, some claim that it could or would change once 'technological singularity' is reached. And of course, like every other term or concept in this discussion, technological singularity means very different things to different people, as illustrated by this article, which provides 17 different definitions: N. Danaylov, 17 Definitions of the Technological Singularity, Institute for Ethics & Emerging Technologies, 22 August 2012, http://ieet.org/index.php/IEET/more/danaylov20120822.

40 This of course includes Asimov's classic "I, Robot" short story series, which developed the three 'laws of robotics' and a host of creative works which followed, including the Terminator movie series, in which Skynet, a computer-run defence network, becomes self-aware and attempts to wipe out the human race; See I. Asimov, I, Robot, 1950.

41 See Report of the 2014 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), http://www.unog.ch/80256EDD006B8954/(%20httpAssets)/350D9ABED1AFA515C1257CF30047A8C7/$file/Report_AdvancedVersion_10June.pdf.

thought it premature to even attempt to find one.[42] As discussed above, trying to precisely define autonomy does not seem possible. And the conversation was only further complicated by the idea that one indefinable concept – autonomy – might be conceived in reference to an almost equally indefinable concept, 'meaningful human control'.[43] This trend proceeded in a CCW LAWS meeting in November 2015, where delegates agreed to continue discussions without resolving any substantive issues, or even reaching consensus as what those issues were. And, as already discussed, in the meetings in April 2016, Human Rights Watch insisted that the conversation was about future systems, while the ICRC reiterated that current systems would have to be included.[44]

There is no reason to believe that this will change at the upcoming LAWS GGE, unless and until the manner changes in which LAWS are considered. The CCW provides a forum for States Parties to discuss certain conventional weapons which are deemed to be excessively injurious or indiscriminate. Framing the LAWS discussion in terms of full autonomy confuses an already complicated area, and introduces non-weapons functions that are outside CCW's weapons treaty purview. Most importantly, full autonomy distracts from CCW's humanitarian focus, banning or restricting the use of *specific* types of weapons that cause unnecessary or unjustifiable suffering to combatants, or indiscriminately affect civilians.

Conclusion

Advances in autonomy will, and indeed already do, herald a technological revolution.[45] Driverless cars, the prospect of package delivery via largely autonomous aerial delivery systems, and infinitely more applications are now being used or developed.

42 Ibid, at para. 17, reporting that "[t]he issue of a definition was raised by a number of delegations". While some suggested that a clarification would be necessary at a certain stage if more substantial work were to be undertaken, most of the delegations indicated that it was too early to engage in such a negotiation.

43 Ibid, at para. 20, reporting that "[m]any interventions stressed that the notion of meaningful human control could be useful to address the question of autonomy".

44 See ICRC (note 3).

45 See P. Singer, The Robotics Revolution, Brookings Institute, 11 December 2012.

The LAWS debate is proving to be a proxy for broader technology[46] and morality questions.[47] And while those questions are fairly asked and debated, in terms of LAWS, the CCW discussions will not benefit, or will never progress, from yet another circular lap mired in trying to define the indefinable, and framed in terms of full autonomy.

The international community needs to leave the cul-de-sac into which the LAWS discussion has driven. Focusing on the critical function of selection and engagement of targets may offer one such possibility. But the most important step is to recognise the circular nature of the discussions to date, which, absent a change in how LAWS are considered, will only continue.

46 B. Appleyard, The New Luddites: Why Former Digital Prophets Are Turning Against Tech, New Republic, 5 September 2014.

47 Statement of the Chairman International Committee of the Red Cross to Convention on Certain Conventional Weapons Meeting of Experts on Lethal Autonomous Weapons Systems, Autonomous Weapon Systems: Is it Morally Acceptable for a Machine to Make Life and Death Decisions?, Geneva, 13-17 April 2015.

Lethal Autonomous Weapons Systems and the Risks of 'Riskless Warfare'

Pablo Kalmanovitz*

Introduction

Some critics of lethal autonomous weapons systems (AWS) hold that such weapons are inherently unlawful under international humanitarian law (IHL), for at least two reasons: (i) they cannot possibly comply with the core IHL principles of discrimination and proportionality, and (ii) they necessarily preclude making human agents fairly accountable for their wrongful effects. Contrary to these critics, I argue in what follows that (i) while IHL would significantly limit the way in which AWS could be permissibly designed and deployed, neither discrimination nor proportionality flatly proscribe their use; and (ii) AWS do not necessarily preclude the fair attribution of responsibility, including criminal liability, in human agents.

Having indicated how AWS may be compatible with IHL and fair accountability, I turn to examine some barriers to effective regulation and briefly indicate ways in which they could be addressed.[1] While IHL-compatible AWS could in principle be developed and deployed, and agents in charge of designing, testing and deploying AWS could be held accountable for the wrongful harms they may cause, states currently have troublingly few incentives to duly control and minimise risks to civilians.

The document is organised as follows. In Section 1, I argue that AWS could be designed and deployed in keeping with the IHL principles of proportionality and distinction; these principles limit permissible deployment, but they do not altogether preclude their use. In Section 2, I argue that AWS deployment need not undermine the basis for fairly attributing responsibility to human agents; neither fair criminal liability nor lesser forms of liability are inherently incompatible with the autonomy of so-called killer robots. Section 3 discusses four barriers to adequate testing and effective compliance with IHL, and concludes.

* Associate Professor, Universidad de los Andes, Bogotá, Colombia. This chapter is based on Pablo Kalmanovitz, Judgment, Liability and the Risks of Riskless Warfare, in N. Bhuta et al. (ed.), Autonomous Weapon Systems: Law, Ethics, Policy, Cambridge 2016, pp. 145-163.

1 While a complete ban of AWS is arguably the best path to take, in what follows I focus on regulation and oversight.

AWS under IHL

The defining mark of AWS is that, once activated, they can "select and engage targets without further intervention by a human operator".[2] The ultimate decision to kill rests on a machine, not on a human operator. Anti-personnel mines count as AWS according to this definition, but the unprecedented and most troubling issues raised by AWS follow from the use of cutting-edge technology to design far more sophisticated and truly autonomous decision-making processes.

Given that weapons technology already performs a wide range of functions autonomously, it may seem as if the automation of the actual use of force would only be another step in an ongoing trend. Critics have argued to the contrary. It has been said that deploying AWS amounts to crossing a 'principled boundary' in a technological slippery slope, namely the principle that "a human being needs to be meaningfully involved in making the decision of whether or not lethal force will actually be used in each case". Crossing this boundary amounts to giving up human responsibility and judgment in the use of lethal force, which is inherently wrong.[3]

There are two distinct worries in this challenge. First, the type of judgments required by IHL is not amenable to computer modelling and cannot be machine-learned. Second, automating the decision to use lethal force would necessarily make it unaccountable.

To address the first worry, consider the principle of proportionality in IHL. As stated in Article 57 of Additional Protocol I, proportionality mandates that "those who plan or decide upon an attack shall refrain from deciding to launch any attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated".[4] As the ICRC Commentary notes, the proportionality rule is in reality an open-ended principle that calls for an "equitable balance between the necessities

2 U.S. Department of Defense, Directive 3000.09, Autonomy in Weapon Systems, 21 November 2012, http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf, p. 13.

3 P. Asaro, On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making, 94 International Review of the Red Cross 687 (2012) 707; M.E. O'Connell, Banning Autonomous Killing, in M. Evangelista and H. Shue (ed.), The American Way of Bombing Ithaca 2014, 224, 232. See similarly N. Sharkey, Saying "No!" to Lethal Autonomous Targeting, 9 Journal of Military Ethics 369 (2010); N. Sharkey, The Evitability of Autonomous Robot Warfare, 94 International Review of the Red Cross 787 (2012).

4 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Additional Protocol I) 1977, 1125 UNTS 3.

of war and humanitarian requirements".[5] Very often – perhaps in most cases – judgments of proportionality will involve hard, indeed incalculable, choices over which people can reasonably disagree.

On one side of the proportionality balance, there is the "incidental loss of civilian life", for which there may be standardised algorithmic estimation methods. The United States armed forces currently apply a 'collateral damage estimate methodology' which quantifies potential damage on the basis of factors such as the precision of a weapon, blast effect, attack tactics and the probability of civilian presence near targets.[6] This procedure could potentially be automated in a machine with enough data on terrain, civilian presence, weaponry, and so on.

However, this estimation method is used only in order to determine the level of command at which an attack that is likely to harm civilians must be authorised—the greater the risk, the more authority required for clearance. The rationale behind this procedure is that the assessment of the "concrete and direct military advantage anticipated", which is on the other side of the proportionality balance, is for a commander to make. The more potential harm there is to protected persons and goods, the more judgment and experience the commander must have when deciding if such harm would be "in excess" of military advantage.

Could this balancing estimation be automated? Possibly yes, but only within narrow settings.[7] Schmitt and Thurnher argue plausibly that AWS could be programmed with collateral damage thresholds under certain limited conditions. They give as an example setting "a base maximum collateral damage level of X for [destroying an enemy] tank". Conceivably, in such cases, a level of maximum allowable side effects could be assigned *ex ante*. Sufficient knowledge would be required about the area where the tank is located and the contribution of its destruction to the war effort. For operations narrowly defined in time and space, threshold values could be set and updated in real time as the operations unfold.

5 C. Pilloud et al., Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949, Leiden 1987 p. 683.

6 M. Schmitt and J. Thurnher, "Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict, 4 Harvard National Security Journal 255 (2013).

7 Most critics of AWS deny that the required balancing estimation could be automated, but this flat denial seems unwarranted. Compare Asaro (n 3); O'Connell (n 3); Sharkey (n 3); and also, if from a different legal angle, E. Lieblich and E. Benvenisti, The Obligation to Exercise Discretion in Warfare: Why Autonomous Weapons Systems Are Unlawful, in N. Bhuta et al. (ed.), Autonomous Weapon Systems: Law, Ethics, Policy, Cambridge 2016, pp. 245-283.

However, to appreciate how wide this window of permissibility is, it is important to be clear on the nature of proportionality judgments – they must be both context-specific and holistic. They are context-specific because both civilian risk and military advantage are highly situational, uncertain, complex and dynamic, and they are holistic because the military advantage of tactical actions has to be assessed relative to broader strategic considerations. As the ICRC Commentary puts it, "an attack carried out in a concerted manner in numerous places can only be judged in its entirety".[8] The tactical value of destroying one particular tank, and, hence, the admissibility of collateral damage, can vary depending on what is happening elsewhere in the war at that point.

The twofold requirement of contextual and holistic determination indicates how inescapable human decision-making is in proportionality estimations. Tactical value is connected to strategy, which is ultimately connected to political goals. It is absurd to think that AWS could make proportionality judgments beyond narrow circumstances, as it amounts to expecting that they could decide on the value of strategic and, ultimately, political goals. The prospects of a machine creatively engaging in discussions of strategy and political goals are remote, certainly not within the foreseeable future of technological development.[9] AWS's algorithmic calculations necessarily belong to narrow tactical actions.

Furthermore, setting threshold values for proportionality assessments in narrow settings would by no means make human judgments superfluous. To the contrary, human judgment remains indispensable. Machines themselves cannot make the type of judgment that proportionality requires, but they could be programmed to proceed on the basis of human-made algorithms and choices of threshold values. Machines may perhaps be capable of attacking with 'surgical precision', but not of balancing reasonably the two sides of the proportionality rule. This impossibility is not technological but conceptual.

Indeed, in the context of AWS, the test of proportionality applies to the *human decision* to deploy AWS in a certain setting and to assign particular parameters of operation. To use the language of the ICRC Commentary, the human operator who deploys AWS would have to assess "in good faith" and according to "common sense" whether the interests of protected persons are duly taken into account in this deployment – given the AWS algorithm and action parameters and the specific conditions of its deployment.

8 Pilloud et al. (n 5); see also Human Rights Watch, Losing Humanity: The Case against Killer Robots, 2012, https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf pp. 32-34.

9 See however A. Krishnan, Killer Robots: Legality and Ethicality of Autonomous Weapons, Farnham 2009, pp. 53-55.

The notion that algorithmic calculations could substitute for reasonableness rests on a misunderstanding of the nature of IHL proportionality. Reasonableness in law refers minimally to an expectation of due care in balancing conflicting values and legitimate interests in complex circumstances. The common element in the various legal uses of the reasonableness criterion "lies in the style of deliberation a person would ideally engage in, and the impartial attention he would give to competing values and evidences in the given concrete setting".[10] Legal procedures involving standards of reasonableness consist of allegations and counter-allegations regarding an agent's proper weighing of conflicting relevant reasons, which the law itself does not settle *ex ante*.

When invoking a reasonableness test, the issue to be decided is whether sufficient reasons exist to act in ways that nonetheless undermine certain legitimate interests or values. Correspondingly, unreasonableness refers to failures in human judgment, most clearly gross distortions in relative valuations, neglect of relevant interests, or the plain failure to bring in relevant reasons.[11] The clearest form of unreasonableness in IHL proportionality is the neglect of fundamental civilian interests for the sake of negligible tactical advantage.

The key point for present purposes is that AWS are not the type of agent that could engage in the practice of giving reasons. By contrast, the human agents who design and operate AWS can – indeed must – give reasons and be held accountable in contexts of practical deliberation, in this case IHL. The relevant question in these contexts is what reasons *they* had to conclude that, given the AWS's algorithm, specifications, and expectable performance, it would not harm the protected interests in excess of what would be gained by its use.

Consider now the IHL principle of distinction. Article 57 of Additional Protocol I mandates that "those who plan or decide upon an attack shall do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects". As the ICRC Commentary notes, the "everything feasible" clause is open-ended and indicative of a margin of judgment; it can be interpreted as mandating decision-makers to take all reasonable steps to identify and avoid harm to protected persons and objects.[12] What counts as feasible or reasonable will vary with the situation and available technology; a balance must be made between the protection of civilians and the advancement of military objectives.

10 N. MacCormick, Reasonableness and Objectivity, 74 Notre Dame Law Review 1575 (1999) 1581.

11 See further O. Corten, The Notion of "Reasonable" in International Law: Legal Discourse, Reason and Contradictions, 48 International and Comparative Law Quarterly 613 (1999).

12 Pilloud et al. (n 5), pp. 678-683.

Schmitt and Thurnher have suggested the test of a 'reasonable attacker' for autonomous weaponry. When estimating whether a potential target is legitimate, it should be considered what a reasonable human attacker would do if he had all of the information available via AWS sensors – if the human attacker would fire, it is legitimate for the AWS to fire.[13] Their test is misplaced in two ways that reveal the practical force of the reasonableness criterion when applied to AWS.

First, the cognitive capacities of a human attacker need not be the baseline for the potentially far superior processing capacity of machines. As defenders of AWS have been keen to emphasise, machines can be programmed to take higher risks in order to acquire more information to accurately identify targets before using force.[14] The risk of self-destruction must weigh differently in humans and machines. Second, the relevant subject for a test of reasonableness is not what the machine should do but rather what the human beings "who plan or decide upon an attack" should do before deploying AWS. Among other things, it must be considered whether *they* have taken all feasible measures to minimise harm to civilians, including sufficient testing and performance estimates. The relevant consideration is not what a human being would do if he were in the AWS's place, but rather how human operators must proceed when – and what they should know before – fielding AWS.

Whether AWS could be fittingly designed for permissible deployment would have to be argued case by case. Sensors that are capable of detecting certain weapons could be presented as a reasonable basis for distinction under some circumstances but not in others. It seems unlikely that in counter-insurgency scenarios, algorithms could perform autonomously and in keeping with the IHL criterion of 'direct participation in hostilities'.[15] Such environments may be too complex and 'cluttered' to allow for automated force. It may well be that compliance with IHL will confine AWS to battle spaces with low or null civilian density, for example submarine warfare, missile shields or swarm technology. But it need not be the case that IHL flatly proscribes AWS.

13 Schmitt and Thurnher (n 6), pp. 262-265.

14 See e.g. K. Anderson and M. Waxman, Law and Ethics for Autonomous Weapons Systems, American University Washington College of Law, Research Paper no. 2013-11 (2013), p. 1; Schmitt and Thurnher (n 6), pp. 248-249; R. Arkin, Governing Lethal Behavior in Autonomous Robots, Farnham 2009, pp. 29–36.

15 N. Melzer, Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law, Geneva 2009.

Accountability for AWS deployment

Critics of AWS have also argued that they should be banned because they inherently preclude the fair attribution of responsibility. Since, by definition, AWS autonomously make the decision of which targets to engage, it would be unfair to hold commanders or anyone liable for a robot's decision. Fair criminal liability presupposes that commanders can foresee and intend the outcome of their actions, both of which are necessarily excluded by the autonomy of 'killer robots'.[16]

This objection to AWS presupposes, first, that machine autonomy necessarily makes the outcome of deployment completely unknown and, second, that accountability is fair only when an agent expressly intended to commit a foreseeable deed. Both are overly strong and unwarranted. Uncertainty and accountability are in fact connected in intuitively obvious ways: deploying a lethal AWS under complete uncertainty would be clearly wrong, comparable to releasing a toxic substance in an inhabited environment. Even if the person did not specifically intend to harm anyone, the harmfulness of the released agent is enough for criminal liability, either criminal negligence or recklessness.

Critics of AWS are of course rightly concerned about giving lethal decision-making powers to robots whose autonomous performance in complex environments would be uncertain. The 'computer revolution of warfare', which includes the extensive use of computerised sensors and complex data analysis, and the prospective use of artificial intelligence for pattern recognition and other tasks, elicit reasonable fears that machines will eventually be created that "can develop behaviors we did not anticipate and that we might not even fully understand".[17]

In civilian contexts, autonomous cars and personal care robots – for example, robots assisting elderly people in retirement homes – already pose great challenges of risk control and public safety. Risk control in robotics depends on systematic testing under controlled conditions. Just as personal care robots and autonomous cars are subject to product liabilities and would have to be approved by public safety agencies before commercial use, AWS would have to be tested by equivalent military and humanitarian agencies. For a military commander to be confident that he has taken all reasonable steps to avoid creating excessive risks to civilians, which he has a legal duty to do, he would have to reasonably trust that sufficient testing has been done on the AWS.

16 Human Rights Watch (n 8), p. 42; likewise Asaro (n 3); O'Connell (n 3); R. Sparrow, Killer robots, 24 Journal of Applied Philosophy 62 (2007).

17 Krishnan (n 9), p. 58.

At a minimum, I submit, sufficient testing should provide the basis for defining a *probability distribution over the set of outcomes in a deployment scenario*; for each possible outcome, commanders should be able to estimate the expected impact on civilians. While AWS may be non-deterministic and non-scripted to various degrees, in all cases their range of action should be bounded and subject to probabilistic estimation. Deploying AWS under second-level uncertainty – uncertainty about their probabilistic range of action – would generate inestimable risks on a civilian population, *and would consequently be unlawful*.

The absolutely fundamental point is that the deployment of AWS itself is not an automated action but the deliberate decision of a human commander. If there is uncertainty – as there will necessarily be in the context of AWS – the question is what steps have been taken to sufficiently limit this uncertainty and sufficiently determine the risks created to civilians. If the range of action and corresponding risk cannot be anticipated with sufficient epistemic confidence, then it would be wrong to field the weapon, possibly a case of criminal negligence. The greater the uncertainty, the less reasonable it would be to believe that the risks are negligible.

The question is of course how much ex ante knowledge is enough – suitable standards and regulations have to be created. But it is already clear that IHL may limit the application of artificial intelligence and machine learning in the context of AWS. To the extent that the outcome of such applications is beyond human control or probabilistic estimation, they would violate the principle that commanders anticipate non-negligible risks to protected persons and goods.

Currently, under international criminal law, superiors who fail to avoid non-negligible risks to protected persons and goods can be liable for negligence or recklessness.[18] This form of liability should be extended to the deployment of AWS.[19] AWS would raise novel and complex issues that require the further development of existing legal frameworks. For instance, IHL standards are currently addressed to the moment of selection of targets, not to the deployment of lethal weapons that themselves undertake more or less open-ended selections. Similarly, the doctrine of command responsibility, which underlies liability for criminal negligence, refers to the relationship between commanders and subordinates, not between commanders and machines. Both legal regimes would have to be extended to cover the novel situations created by AWS.

18 K. Ambos, Superior responsibility, in A. Cassese et al. (ed.) The Rome Statute of the International Criminal Court: A Commentary, Oxford 2002, p. 805.

19 This natural legal extension is defended by N. Jain, Autonomous Weapons Systems: New Frameworks for Individual Responsibility, in N. Bhuta et al. (n 7), pp. 303-324, and G. Corn, Autonomous Weapons Systems: Managing the Inevitability of 'Taking the Man Out of the Loop', in N. Bhuta et al. (n 7), pp. 209-242.

Criminal sanctions against commanders would be the most extreme form of liability, but other forms in other agents are conceivable as well. A large and complex command structure is implicated in the operation of new weapons technologies. Those who actually field AWS would have to assume that their superiors and legal experts had duly overseen that the designers, programmers and testers had done their jobs competently.

Criminal liability aside, wrongful harm to protected persons and goods would have to be felt through the complex structure of the deploying of a state's armed forces, possibly in relatively milder forms of liability such as disciplinary measures, professional disqualification and civil remedies. While developers, engineers, lawyers and technicians may be far from battlefields and, as such, not directly responsible for AWS fielding, they could be professionally disqualified for their failure to anticipate shortcomings or to duly test for risks.

Barriers to effective control of AWS

I have argued that, in virtue of IHL, it is mandatory to test and determine the risks generated by AWS, and that human agents may be fairly held accountable for wrongful harms caused by fielding. A fundamental objection to AWS has thus proven to be unfounded. However, my discussion so far has taken place at an idealised level. While a regime of due care in testing and risk control, with suitable rules of individual liability, is in principle possible, the question of whether such a regime is *likely* to be implemented, and whether states developing AWS would have *enough incentives* to fund and implement demanding risk control programs, and make their own officers, lawyers, and contractors accountable, is a different matter.

AWS will add to the already stark asymmetries of so-called riskless warfare, in which risks are increasingly shifted to foreign civilians.[20] The aversion of states to endure casualties among their own forces and their reluctance to deploy 'boots on the ground', together with the imperative to maximise 'manpower efficiencies', feed the development of means of force that generate and externalise virtually all risks to enemy combatants and civilians. High-altitude bombings in Kosovo and elsewhere have already illustrated this trend, but drone technologies have deepened it, and AWS would do so even further. It is to be expect-

20 P. Kahn, The Paradox of Riskless Warfare, 22 Philosophy and Public Policy Quarterly 2 (2002); D. Luban, Risk Taking and Force Protection, in Y. Benbaji and N. Sussman (ed), Reading Walzer, Abingdon-on-Thames 2014, pp. 277–301.

ed that the few states capable of developing and deploying AWS may further shift risks to foreign civilians, and it is by no means clear how this increased endangerment could be effectively reversed or constrained.

In contrast to robotic technologies designed for civilian uses – for which national consumer protection agencies exist which set standards of risk and testing – for lethal autonomous robots there are no international agencies designated to protect civilians against the risks generated by AWS. Article 36 of Additional Protocol I mandates that new weapons be duly tested for compatibility with IHL, but it does not institutionalise this duty at the international level. State Parties explicitly declined to set up international bodies to oversee the reviewing process.[21] Each AWS-developing state is in charge of testing, which raises several concerns.

First, states are of course interested parties in the review process, as one of the stronger motivations for designing and deploying AWS is precisely to minimise risks to their own armed forces. Foreign civilians have no voice or representatives in the review process. Furthermore, foreign civilians may either be too remote or held too much in suspicion to elicit sympathy among the deploying states' politicians or citizenry, or any potential sympathy may be obstructed by official secrecy and obfuscation.

Second, the testing and evaluation of new weaponry is time-consuming and expensive, both of which run against the budgetary constraints, tactical interests, and sense of urgency of deploying states. In times of shrinking defence budgets, it will be very tempting to opt for expediency. A recent commentary on AWS testing lays down the relevant trade-off clearly, but revealingly does not rule out deployment even if testing is manifestly insufficient: "[I]f a high statistical assurance is deemed necessary for civilian safety while budgetary constraints preclude the corresponding necessary development testing, then appropriate limits should be implemented regarding the approved applications for that weapon until field experience provides appropriate reliability confidence".[22] The risks of such a 'test-as-we-go' approach are obvious, as the point of testing is precisely to determine *ex ante* if fielding would be overly harmful to protected goods and persons.

Furthermore, not only would each AWS have to be tested before fielding, but whole new testing methods would have to be created. According to a recent report on AWS by the US Defense Science Board, performance benchmarks in testing and evaluation are usually

21 Pilloud et al. (n 5), pp. 421-428.

22 A. Backstrom and I. Henderson, New capabilities in warfare, 94 International Review of the Red Cross 483 (2012) 509.

pre-scripted and deterministic, while AWS may require a 'systems engineering approach' with yet undefined probabilistic criteria.[23] Not only do AWS perform non-deterministically, they raise unprecedented and costly challenges to testing design and execution.

Third, it was argued above that references to reasonableness in the principles of distinction and proportionality make human judgment indispensable in AWS deployment. The flip side of this is that standards applicable to AWS deployment are vague and open-ended, and subject to change with technological innovation.

References to reasonableness in law often have the function of concealing diverging interpretations and unresolved political tensions. Historically, this was certainly the case as regards IHL proportionality, as states opted for an open-ended rule after disagreeing on more exact formulations of the proper balancing of civilian protections and military necessity. Such disagreements in international law may often be solved not by the normative force of the most appropriate considerations, but rather by the sheer power of the parties: "each State maintains its own conception of what is reasonable, and will exercise its powers according to that conception".[24] Sheer power, rather than the best argument (regarding e.g. maximum protection of civilian interests), carries the day, which means that the interest of the stronger will prevail against the interests of foreign civilians who lack political voice or power.

In the context of AWS, it is to be expected that the stronger party will have AWS capabilities, and so the open-ended reasonableness criterion is likely to incorporate its preference for shifting risks away from its own forces and enjoying economies in testing. The principles of proportionality and sufficient precautions will gradually shift to reflect this preference. Higher damage to civilians will come to be treated as proportional; uncertainty will be found admissible if not harmful to one's own forces. Tactical gains may prove simply too alluring when testing is very expensive and time-consuming, and humanitarian failure has no tangible costs.

23 U.S. Department of Defense, Defense Science Board, Task Force Report: The Role of Autonomy in DoD Systems, 2012, pp. 62-64.

24 Corten (n 11), pp. 618-620.

Fourth and last, it may be hard to attribute responsibility for AWS deployment to specific individuals in the complex web of industrial-military personnel behind their production, testing, and deployment. Responsibility will dissolve to a vanishing point in that network unless states are willing to conventionally agree to define certain roles as responsible for the consequences of AWS deployment.

Conclusion

To conclude, these four barriers do not amount to saying that states equipped with AWS will be completely unconstrained in their design and deployment. They may be motivated by the legitimacy costs of perceived disproportionate harm to civilians. Public opinion and the political consequences of naming and shaming may be a very real concern to decision makers, and, at least to this extent, states may have an interest in being perceived as law-compliant, however elastic legal criteria may be. The use of easily accessible technologies to record the impact of AWS and to broadcast it in social media may have real disciplining effects. There is also the imperative to minimise risks to friendly forces, which may require considerable amounts of testing and safety reviews, as well as the ethos of the military profession, which may set limits on the range of actions military men are willing to undertake with AWS. These are all relevant and potentially effective constraints.

However, unless the further erosion of IHL principles is prevented, and unless suitable extensions of IHL and international criminal law are agreed, in particular regarding binding standards on testing, it is to be feared that these barriers to effective control will prove insurmountable. AWS would then have utterly unforeseeable but surely disastrous effects on civilian lives and wellbeing.

Mapping Autonomy

Leon Kester*

Introduction: Applications for Autonomy

Autonomous systems play an ever-increasing role in modern society. Typical applications where ethical issues play a role are mobility and defence.

Regarding mobility, autonomous vehicles may have to make decisions on how much risk to take, how close to drive to other vehicles or whose life to spare and whose life to put in danger. In the defence domain, much attention is on the problem of autonomous vehicles as weapons systems making decisions on life and death.

In order to address these issues for various domains, in this section, a generic approach is taken. First, we map autonomy in automation and learning, and for this type of autonomy consider the role of the human. Then, we look at a third aspect, self-adaptation, and reconsider the role of the human. Finally, considerations on ethics of such autonomous systems are discussed.

Autonomy: Automation

The basis for autonomy is the automation of situation assessment and situation management.

Automatic situation assessment may consist of sensing with various dedicated sensors, processing the signals of theses sensors, detecting interesting objects, following those objects in time, classifying the objects, identifying the objects, finding relations between the objects, and determining the possible risks or benefits those objects may cause. For objects that have a form of intelligence, assessment of the intent of the object is also important.

Situation assessment can thus be considered a chain of processes generating representations of the situation at an ever-increasing abstraction level. The lower levels are characterised by simple representations needing large amounts of data, while at the higher levels, the algorithms and representations are more complex but less data intensive, since only important information needs to be maintained.

* Senior Research Scientist at TNO, The Hague.

Current developments of algorithms for situation assessment are focusing on extension from the lower levels, that means signal processing, detection, and following the objects towards the higher levels. Also on the higher levels, the representations need to be more in line with the semantic representations that humans use to characterise the situation. For the application, domains like mobility and defence and also the situations are not only complex but highly dynamic, so the range of possible situations the autonomous system needs to assess is huge.

Automatic situation management uses the results of situation assessment, which is also called situation awareness, to manage the situation. Automatic situation management faces the same challenges as situation assessment, that means it becomes more difficult on the higher levels due to the increased complexity of the situation, hence requiring more complex representations. Therefore, algorithms are more tedious to develop and use more processing power and memory for background knowledge of the world.

Another trend in automation is the cooperation between multiple autonomous systems with situation assessment and/or situation management capabilities. Cooperation is easier on the higher levels of situation assessment and situation management, since only small amounts of data need to be exchanged, so that the cost of communication is acceptable.

However, cooperation on the lower levels will result in more consistent and higher quality situation assessment and management. Since communication costs are getting lower and filters for selecting the most important information are getting better, cooperation for situation assessment and situation management on lower levels of abstraction becomes more feasible.

Some of the advantages of automation are that it is relatively fast and transparent, meaning that the working of the algorithms can be analysed in order to understand their behaviour. On the other hand, one disadvantage is that it is not dealing well with failures of the system and unexpected situations.

Learning

Another aspect of autonomy is learning. Learning autonomous systems basically use a trial and error method to assess and manage a given situation. The procedures that generate a positive result are reinforced, meaning that they will be favoured next time the same situation occurs.

We can make a distinction between online learning (i.e. learning on the job) and offline learning (i.e. training).

There are many different types of learning, for example evolutionary algorithms, reinforcement learning, ant colony optimisation, deep learning, and so on. It is beyond the scope of this paper to discuss the developments and advantages and disadvantages of these methods individually.

In general, developments in learning are particularly focused on making learning methods faster and therefore suitable for situation assessment and situation management in more complex and dynamic situations. Learning will be faster because computers and communication becomes faster, but also because there is a development towards more efficient algorithms. The advantage of learning is that it can be a powerful tool in case the situation is too complex to be modelled comprehensively.

There are a number of disadvantages, however. For real-time applications, learning methods are still very slow. They are also not transparent, meaning that it is difficult to understand why and how they work. Online learning can be unpredictable and is prone to deception. Offline learning, on the other hand, has the same disadvantage as automation in that it does not deal well with failures of the system and unexpected situations in which the autonomous system had no training.

Role of the human

With the above outlined developments of automation and learning in mind, we can distinguish three different roles of humans with respect to (semi-)autonomous systems.

The human cooperates with the autonomous system. Both play their part in situation assessment and situation management. Since autonomous systems still have difficulties in engaging in situation assessment and situation management on the higher levels, those aspects are usually carried out by the human. Aside from the situation assessment and situation management capability of the autonomous system, there is also a need for the human to understand how the autonomous system works. Since automated and learning systems have difficulty in explaining how they work, cooperation between the human and the autonomous systems is a challenge.

The human can also act as a failure-safety agent. This means that the human intervenes when there are failures in the autonomous system, or when the autonomous system is in an unexpected situation, such that erroneous behaviour is to be expected. The third role of the human, furthermore, is that of a moral agent. The autonomous system is not expected to be able – or is not allowed to make – moral decisions. The human therefore makes moral decisions or intervenes whenever the autonomous system is expected to behave immorally.

Self-adaptation

A third aspect of autonomy, aside from automation and learning, that is often not addressed well, is self-adaptation.

First of all, self-adaptation involves self-assessment, that means the processing of the following questions: in what kind of situation am I? What is my goal given this particular situation? What is the current state of my resources: sensors, computation, communication, actuators? How well do I perform situation assessment and situation management? How useful is my information to others? How useful would information of others be for me? How do I appear to the outside world? How certain am I of what I am doing?

Based on this self-assessment, the autonomous system can then manage itself. There is the possibility for it to reason: now that I know and understand myself better, how should I manage myself in order to reach my goals?

Role of the human revisited

When we consider that autonomous systems are capable of adapting themselves to changing goals, changing situations, and changes in their own capabilities, the perspective on the role of the human also changes.

When the system has self-assessment and self-management capabilities, failures of the system can be recognised and managed by the system itself. It is thus no longer exclusively the role of the human to manage all parts of the autonomous system.

Furthermore, the cooperation between human and autonomous system is improved because the autonomous system is now able to explain what it is doing or trying to do, while the human can express his or her needs to the autonomous system. These needs become the goals of the autonomous system and can lead to an adaptation of the system's behaviour.

This also opens up possibilities for autonomous systems to take part in moral reasoning. Not all specifications of the autonomous system are determined during the design phase, in the form of norms. Rather, the autonomous system can derive from higher-level goals what would be desirable in a given situation, during operation. Thus, human and autonomous systems now also have a common ground to cooperate in moral reasoning.

Consideration on ethics of autonomy

The goals that autonomous systems pursue, including goals of an ethical character, should be consistent and explicit. If there is a possibility that the goals cannot be reached, the utility of less desirable outcomes should also be quantified. In this case, the autonomous system can always try to get as close to the desired goal as possible.

It is often argued that the goal (or utility) function is difficult to specify in this way. There are several reasons why this might be less difficult than presumed. If an autonomous system has a self-assessment and self-management capability, it can derive utility from the high-level goal function during operation, so that it does not have to be specified comprehensively during the design phase. There is a reluctance to be too specific about ethical goals, so the difficulty of specifying the utility function is sometimes used as an excuse for not having to specify it.

From the perspective of mission effectiveness of humans and autonomous systems working together, not only the current shortcomings and limitations of autonomous systems should be considered. Rather, one should just as much focus on the shortcomings and limitations of humans, such as cognitive constraints, cognitive biases, or cultural differences.

As a final, particularly relevant remark, it should be pointed out that much attention is currently given to autonomous weapons systems that could make decisions about life and death. It is tempting to ban, for that reason, the development of such systems in order to ease our fears. However, fear is usually a bad advisor. It will draw the attention away from a more difficult question: what if autonomous systems are better in making such decisions – would it not be unethical not to use them? Furthermore, banning would create a false sense of safety, since it is not difficult to separate the development of powerful autonomy from that of weapons systems – and then combine those separate elements just before employment. For that reason, it would be wise to put more effort on world-wide safe development of ethical autonomous systems, regardless of their subsequent application.

LAWS in the Maritime Domain: An Asia-Pacific Scenario

Collin Swee Lean Koh*

Introduction

The shift in the world's economic weight towards the Asia-Pacific region has created a confluence of contradicting trends. It has led to a growing affluence in what is now a buzzing region with bright socioeconomic development prospects underpinned by seaborne trade and commerce. Yet the Asia-Pacific remains beset with a myriad of security challenges, including interstate disputes and unconventional threats in the maritime domain, not to mention an ongoing spate of military build-ups.

Against this backdrop, this section seeks to examine the prospects and challenges of the proliferation of Lethal Autonomous Weapons Systems (LAWS) and their use in the Asia-Pacific maritime context.

Setting the Asia-Pacific Maritime Context

The Asia-Pacific region is not monolithic; it comprises several sub-regions, namely, South Asia (along the Indian Ocean Rim, excluding East African littorals, including Bangladesh, India, Pakistan, Sri Lanka), Southeast Asia (dominated by the ten member states of the Association of Southeast Asian Nations or ASEAN), and Northeast Asia (China, Japan, the two Koreas, Taiwan and the Russian Far East). For the purpose of this study, parts of Oceania (Australia and New Zealand) are also included.

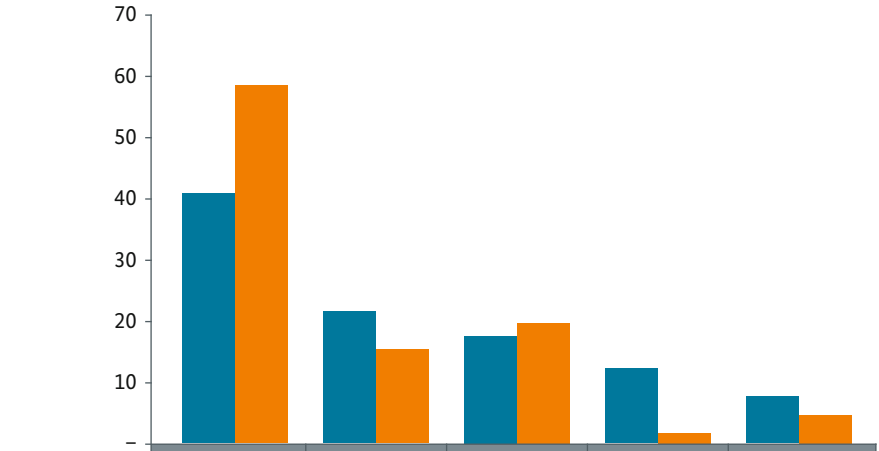
These diverse and disparate sub-regions in turn comprise constituent countries that confront their own national contexts concerning threat perceptions, needs, and capacities. Notwithstanding these differences and disparities, it is the maritime commons that bind the Asia-Pacific littorals. This region straddles both the Indian and Pacific Oceans, charac-

* Research Fellow, Maritime Security Programme, Institute of Defence and Strategic Studies, S. Rajaratnam School of International Studies (RSIS), Singapore.

terised by partly semi-enclosed littoral geography. Many Asia-Pacific states possess long coastlines and vast maritime zones under their jurisdiction. A number of them are sprawling archipelagos made up of countless islands.

But the Asia-Pacific maritime domain presents both a boon and a bane. It contains rich marine hydrocarbons, minerals, and fishery resources which are far from being fully tapped. Some of the world's vital sea lines of communications also pass through the region, for example the strategic Malacca Strait. The region occupies a major share of the world's seaborne trade (see Figure 1).

Figure 1: World Seaborne Trade by Region, 2014 (% Share in World Tonnage)



	Asia	Americas	Europe	Oceania	Africa
Loaded	41	22	18	12	8
Unloaded	58	16	20	2	5

Source: United Nations Conference on Trade and Development (UNCTAD), *Review of Maritime Transport 2015* (Geneva: United Nations, 2015), p. 13.

However, the Asia-Pacific waters are fraught with a multitude of security challenges, such as unresolved maritime disputes – for instance the simmering tensions in the East and South China Seas as of late – and persistent transnational threats, for example maritime terrorism, illegal fishing, and smuggling. In an era of global economic interdependence and deep connectivity, these security problems potentially have a dire global impact.

The significant economic and strategic stakes, seen from the perspective of the complex and uncertain Asia-Pacific security environment, motivate the spate of regional military modernisation efforts. A survey of the military equipment procured post-Cold War illustrates preference in force projection capabilities within the maritime domain – large surface warships, submarines, multi-role fighter jets and various types of anti-ship and land attack cruise missiles. Altogether, they generate enhanced levels of mobility, precision, firepower, and protection.

Most of this weaponry consists of manned platforms and, not counting supervised autonomous systems (e.g. fire-and-forget missiles), a growing proportion of semi-autonomous (commonly called 'unmanned') aerial, surface and sub-surface assets. Clearly therefore, these autonomous weapons systems – both lethal (used for combat) and non-lethal (used for surveillance and other purposes such as mine-clearance), are already used by Asia-Pacific militaries.[1] Of key interest is the future role of increasingly autonomous weapons systems in the region's maritime domain.

1 A semi-autonomous system performs a task and then stops and waits for the human operator's approval before continuing, hence 'human in the loop'. Unmanned aerial, surface, and sub-surface systems would fall under this category. A supervised autonomous system, once activated, performs a task under human supervision and will continue performing the task unless the human operator intervenes to halt its operation, hence 'human on the loop'. Generally, advanced fire-and-forget weapons fall under this category. Fully-autonomous weapons, however, once activated, perform a task and the human operator does not have the ability to supervise their operation and intervene in the event of system failure, hence 'human out of the loop'. Such a weapon does not exist yet. See P. Scharre, Autonomous Weapons and Operational Risk, Ethical Autonomy Project, Center for a New American Security, Washington, DC 2016, pp. 9-10.

LAWS: more problems than appeal?

To better understand LAWS proliferation in the Asia-Pacific maritime context, it is important to briefly examine this particular weapon system's appeal and problems. Weapon automation, which is inevitable as a response to the increasing tempo of military operations and political pressures to protect not just one's own personnel but also civilian persons and property,[2] logically heralds increasingly autonomous systems. This process is driven by the following aspects:[3]

- greater force projection through more efficient use of fiscal and manpower resources than traditional troops and man-in-the-loop robots in the face of dwindling defence budgets and high personnel costs;
- ability to act without regard to self-interest, thereby reducing exposure of and risk to humans in executing 'dirty and dangerous' missions;
- freeing humans from dull or repetitive tasks;
- less susceptible to certain forms of remote intervention by unauthorised entities;
- access to larger amount of information, increased speed of decision-making, shorter response time from identification of potential targets to implementation of a planned reaction (i.e. compressed response window), greater accuracy and greater predictability for certain functions in specific environments; and
- a belief by some that such systems may eventually be able to respect international humanitarian law or human rights law better than humans do.

The unique factors of the maritime environment also drive increasing autonomy.[4] But despite these benefits, autonomous technologies carry inherent risks as a function of the task being performed and its operating environment.[5] Fully-autonomous technologies

2 K. Anderson and M.C. Waxman, Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can, Columbia Public Law Research Paper No. 2013-11, Washington, DC 2013, p. 2.

3 United Nations Institute for Disarmament Research, Framing Discussions on the Weaponization of Increasingly Autonomous Technologies, Geneva 2014, pp. 5-6; G. Bills, LAWS unto Themselves: Controlling the Development and Use of Lethal Autonomous Weapons Systems, 83 The George Washington Law Review 176 (2014).

4 These factors are: environmental and economic factors; limits of human personnel especially in the conduct of extended submarine missions; difficulty of maintaining communication with systems underwater; and benefits obtained from the covert nature of such a communication-denied environment as the underwater realm; see United Nations Institute for Disarmament Research, The Weaponization of Increasingly Autonomous Technologies in the Maritime Environment: Testing the Waters, Geneva 2015, p. 2.

5 Scharre (n 1), pp. 9-10.

are significantly more complex compared to semi- and supervised autonomous systems. As such, it becomes increasingly difficult for a human operator (even for well-trained ones) to predict precisely what the autonomous system might do in any given situation.[6]

While the behaviour of autonomous systems may be predictable and safe under most operating conditions, the inability to confidently predict the behaviour of the system under all possible operating conditions raises uncertainty regarding the conditions under which the system will fail. This makes it more difficult for the human operators to avoid failures. When combined with autonomous systems that have a high damage potential if they fail, the result could be significant risk, especially when the weapon selects and engages targets other than what the human operator intended, resulting in fratricide, civilian casualties, or unintended escalation in a crisis.[7]

Moreover, by reducing the risk to human operators, LAWS may also reduce the political costs and thereby lower the threshold of going to war. Since LAWS is most plausibly going to be restricted to just a small handful of advanced countries at the onset of their proliferation, the systems' entry in particular into a volatile geopolitical environment may become a source of destabilisation.

Destabilising weapons may be deemed to have fulfilled all or some of the following qualities: result in decreased warning time; give one country 'breakthrough' capabilities; lead to a broadening of target sets; permit no effective countermeasures; give one side better information concerning another's military preparations; and create hostility.[8] Especially at the initial stages of their introduction into the global weapons arena, LAWS may fulfil some of these attributes, especially the part about 'breakthrough' capabilities and altering the balance of military power, engendering interstate hostility especially in a geopolitically volatile region. Therefore, LAWS may create instability and increase the likelihood of armed conflict.[9]

6 Ibid, p. 12.

7 Ibid, pp. 17-18.

8 D. Mussington and J. Sislin, Defining Destabilizing Arms Acquisitions, Jane's Intelligence Review, Vol. 17, No. 2 (February 1995), pp. 88-90. See also C.A. Meconis and M.D. Wallace, East Asian Naval Weapons Acquisitions in the 1990s: Causes, Consequences, and Responses, Westport 2000, pp. 35-36.

9 P. Asaro, On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decisionmaking, 94 International Review of the Red Cross 692 (2012).

Problems of LAWS in the maritime domain

To better illustrate the risk of destabilisation caused by LAWS in the maritime domain, it might be helpful to examine the effects of contemporary naval armaments. Modern warships are increasingly sophisticated, versatile and yet accordingly expensive.[10] This means that such platforms are acquired in smaller numbers; each of them designed to fulfil an even greater range of missions. This also makes them very vulnerable to attack; their loss in action may mean a huge blow to operational capabilities and national prestige. Therefore, modern warships are also known as 'one hit ships' – either completely put out of action or sunk in just a single attack.[11]

Modern post-1945 examples include the British destroyer *Sheffield*, sunk by an Argentine air-launched missile in 1982. The Israeli corvette *Hanit* was withdrawn from action after being seriously damaged by a Hezbollah missile off of the coast of Lebanon in 2006. Finally, a more recent case was the South Korean corvette *Cheonan*, which was allegedly sunk by a North Korean torpedo in 2010. The vulnerability of modern warships thus potentially creates a 'use them and/or lose them' dilemma for on-scene commanders: either respond decisively to any tell-tale signs of an impending attack, or risk losing the platform once the attack has been committed. This is especially so for surface warships which are considered extremely 'time-critical systems' associated with high 'first-shot premiums'.[12]

Therefore, the impulse for reflexive action in response to signs of impending attack, such as electromagnetic emissions from the adversary's fire control radar, may provoke accidental clashes.[13] Modern warships have become increasingly vulnerable to long-range missiles backed by sophisticated sensors. Furthermore, shipboard defences may come at disproportionate costs for already expensive warships, thereby disadvantaging their race against cheaper, offensively-oriented anti-ship weaponry. This exerts pressure upon or incentivises the

10 The capital costs of naval programs are so high partly because they involve much more than just acquisition of ships. Shore facilities, logistic support and training personnel require a major investment of human and material resources; see I. Anthony, The Naval Arms Trade, SIPRI Strategic Issue Papers, New York 1990, p. 165.

11 H.J. Kearsley, Maritime Power and the Twenty-First Century, Aldershot 1992, pp. 30-31.

12 M. Chalmers et al., Alternative Conventional Defense Structures for Europe: British and Danish Perspectives, AFES-PRESS Report No. 22, 1989, p. 19; R. Neild, The Implications of the Increased Accuracy of Non-Nuclear Weapons, in A. Boserup and R. Neild (ed.), The Foundations of Defensive Defence, Basingstoke 1990, pp. 61-62.

13 It was argued that the moment when a potential attacker's fire control radar is 'locked on' to the target, this qualifies as an 'armed attack' so that use of force in self-defence may be undertaken; see D.P. O'Connell, The Influence of Law on Sea Power, Manchester 1975, p. 82.

naval combatant to shoot first, thereby escalating the risk of accidental or inadvertent clash.[14] Incidents involving submarines may magnify pressures on the crew to take drastic action in times of tense confrontation, since, compared to surface warships, they stand a higher chance of being lost with all hands on board.[15] Moreover, the need to operate under strict emission control (especially concerning communications with the outside world) constitutes a key operational requirement of underwater operations, thus amplifying those risks.

Besides deliberate action on the part of the human operator, the risks of systems failure, miscommunication and human error may constitute a recipe for disaster. The U.S. Navy frigate *Stark* failed to intercept the incoming Iraqi missiles because its Phalanx close-in weapons system was faulty.[16] Another example was the shoot-down of an Iranian airliner by U.S. Navy cruiser *Vincennes* in July 1988, when in the heat of battle with Iranian gunboats in the Persian Gulf, the crew mistook it for an Iranian fighter-bomber. Finally, the U.S. Navy aircraft carrier *Saratoga* accidentally fired anti-aircraft missiles which struck the Turkish destroyer *Muavenet* during an exercise in late 1992, killing a few on board.

If automated, semi- and supervised autonomous naval weapons can potentially destabilise and even fail to work as intended by the human operator, what can be said of naval LAWS in the future? Given that prior to the commencement of armed conflict, the first unlawful use of force against a state's warships and military aircraft may be perceived as an armed attack on that state, thereby allowing it to invoke the right of national self-defence, what will happen if the warship or military aircraft was unmanned?[17] There are certainly more complications in such complex environment as the maritime domain, for example the political and legal implications of an unmanned attack of an unknown source on a foreign

14 S. Lodgaard and J.P. Holdren, Chapter 1: Naval Arms Control, in S. Lodgaard (ed.), Naval Arms Control, Oslo and London 1990, p. 12.

15 Consider not just submarine losses back during the world wars and peacetime accidents such as the losses of U.S. nuclear submarines Thresher (1963) and Scorpion (1968). The Russian submarine Kursk, which sank in 2000, saw most of its crew killed shortly after disaster struck, whereas the few survivors eventually died of asphyxiation before help arrived.

16 The weapon was found later to have intermittently failed Systems Operability Test number five due to an improperly connected wire in the elevation resolver circuitry, a problem which was unrectified even as the frigate got underway on May 17, 1987; see Formal Investigation into the Circumstances Surrounding the Attack on the USS Stark (FFG 31) on 17 May 1987, Office of the Chairman, The Joint Chiefs of Staff, Department of Defense, 3 September 1987, http://www.jag.navy.mil/library/investigations/USS%20STARK%20BASIC.pdf.

17 A. Backstrom and I. Henderson, New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews, 94 International Review of the Red Cross 497 (2012).

warship – essentially a floating piece of another country's sovereignty – in international waters. For these reasons, the prospects of LAWS proliferation in the Asia-Pacific littorals are a worthwhile topic of examination.

Prospects of LAWS acquisitions in the Asia-Pacific

The Asia-Pacific's diverse and disparate nature means that this vast and complex region cannot be analysed as a whole regarding the prospects of LAWS proliferation. Even within sub-regions, it is impossible to identify common drivers behind each country's potential quest for such a technology. For the purpose of this chapter, the following set of four influencing factors may be derived. Semi-autonomous systems are used here as a point of reference in extrapolating the scenario of LAWS proliferation and use in the Asia-Pacific.

Maritime security threat perceptions

The four different Asia-Pacific sub-regions analysed here present different perceptions of maritime security threats.

Northeast Asia's maritime domain is fraught with mainly traditional security threats emanating from unresolved geopolitical flashpoints: the East China Sea dispute, the Korean Peninsula, Dokdo Islands/Takeshima, Ieodo/Socotra Rock, Taiwan Strait and Kuril Islands. Non-traditional maritime security risks are comparatively lower.

While Southeast Asia's maritime domain is fraught with both non-traditional and traditional security threats, it is safe to say that the former category is predominant if one surveys the various countries' preoccupation with transnational maritime security challenges, such as illegal fishing, piracy and sea robbery, maritime terrorism, smuggling, and human trafficking. There are also unresolved interstate maritime disputes, although many have already been settled peacefully in the 1990s and early-2000s. The key traditional maritime security challenge lies in the South China Sea.

South Asia's maritime domain represents both traditional and non-traditional security challenges. First, there is the longstanding Indo-Pakistani naval rivalry, with an increasing Sino-Indian dimension, in the Indian Ocean. However, non-traditional maritime security issues, including piracy and sea robbery, illegal fishing, and human trafficking, continue to persist. Maritime terrorism also plays a significant part, as it overlaps with the existing

Indo-Pakistani animosity, as exemplified by the November 2008 terror attacks committed by armed militants who entered Mumbai by sea, which resulted in a brief, tense border standoff between Indian and Pakistani militaries.

Oceania is more peaceful compared to the other sub-regions. Maritime security concerns emanate largely from non-traditional sources such as illegal fishing, and, especially in the case of Australia, the influx of asylum-seeking refugee boats, an issue closely linked to the issue of human trafficking faced by countries such as Bangladesh and Sri Lanka.

Military employment of semi-autonomous systems

Semi-autonomous systems offer a useful point to extrapolate militaries' future inclination towards LAWS. Generally speaking, semi-autonomous systems have found service in all Asia-Pacific sub-regions – primarily unmanned aerial vehicles and unmanned underwater vehicles (UUVs) optimised for non-lethal roles such as mine-countermeasures and hydrographic surveys. The key disparity lies in the extent to which they are employed in those respective militaries as a result of costs and manpower capacity.

All Northeast Asian countries operate a range of semi-autonomous systems, mostly short-range tactical and medium-altitude, long-endurance (MALE) drones as well as non-lethal UUVs. But there is also a drive towards high-altitude, long-endurance (HALE) drones. For example, China has developed the Xiang Long (Soar Dragon) – analogous to the American RQ-4 Global Hawk, which recently was acquired by Japan and South Korea. But some Northeast Asian militaries have gone a step further by operating armed UAVs. China, for example, has developed the Cai Hong (Rainbow) series MALE drones – analogous to the American MQ-9 Reaper – capable of launching small tactical missiles. Notably, South Korea operates the Israeli-made Harpy anti-radar drone, which is capable of autonomously loitering over the target area to search and destroy electromagnetic emitters. The Harpy is probably the closest to LAWS.

In Southeast Asia, the proliferation of semi-autonomous systems is less widespread. Only Indonesia, Malaysia, the Philippines, Singapore, Thailand, and Vietnam constitute major operators. Amongst them, Singapore can be considered having the most comprehensive suite comprising UAVs, unmanned surface vessels (USVs), and autonomous underwater

vehicles (AUVs).[18] UAVs in service belong to short-range tactical and MALE types; none of them armed. Some countries, such as Vietnam, have expressed interest in acquiring HALE drones.

This lower extent of semi-autonomous systems employment within Southeast Asian militaries can be attributed to cost and manpower. First, sophisticated semi-autonomous systems are not cheap, especially if acquired from foreign vendors. The second, arguably the most crucial, reason may have to do with the manpower capacities of some of these militaries to absorb such technologies. In fact, the personnel of some Southeast Asian militaries are still trying to grapple with manned weapons systems.[19]

The costs and manpower requirements to operate and maintain such technologies resulted in low levels of proliferation in South Asia. India and Pakistan constitute primary semi-autonomous systems military operators, mainly of Chinese- and Israeli-made short-range tactical and MALE drones. Islamabad has also tested an indigenous UAV armed with missiles, likely developed with Chinese assistance.[20]

Australia is the key operator of semi-autonomous systems in Oceania, maintaining a fleet of short-range tactical and MALE drones as well as UUVs for low-intensity maritime operations. It does not operate any lethal unmanned systems. However, in more recent years, Canberra has been keen to acquire the MQ-4 Triton (Global Hawk variant optimised for broad-area maritime surveillance) and Reaper drone.[21]

Access to technologies

Besides affordability and manpower constraints, securing access to military technologies can be a problem. Foreign imports and indigenous development are two possible routes. Typically, only the more advanced countries are in a position to exercise the latter option.

18 The Singapore Armed Forces operate both Israeli and locally-developed USVs armed with weapons for various tasks.

19 An example is Malaysia, which has one of Southeast Asia's most advanced militaries. But even the Malaysian Armed Forces is still struggling with optimally operating and maintaining its manned assets, not to mention unmanned systems. Author's interview with Malaysian defence analyst, Kuala Lumpur, 2 March 2016.

20 "Armed Drone, Laser-Guided Missile Tested", Dawn, 14 March 2015.

21 Australia's interest in acquiring Reaper drones in the future was illustrated by the commencement of training for Defence Force personnel in February 2015: "Australia Military Training with U.S. Reaper Drones", Dow Jones Institutional News, 23 February 2015.

For the purpose of this section, it is assumed that countries which enjoy secure foreign access or possess domestic research and development capacities are better placed to acquire LAWS. Seen in this light, not many Asia-Pacific countries enjoy reliable access to foreign technologies.

For example, augmenting their own domestic technological bases, Australia, Japan, and South Korea may tap on their bilateral alliances with the United States to secure state-of-the-art American military technologies such as HALE drones, and plausibly LAWS in the future. The Philippines and Thailand also maintain military alliances with Washington, but their access to such technologies is more likely to be tampered by problems of how to afford and absorb them. For many Asia-Pacific countries – especially those which have endured Western arms embargoes – the indigenous pathway is preferred. Yet indigenous development is uneven throughout the region.

In Northeast Asia, China and Russia are key players; seeking to establish a range of such systems for operations in the aerial, surface and sub-surface dimensions. Japan, South Korea and Taiwan only seriously entered the game in recent years, though they may tap their defence and security relationships with the United States to secure easier access. As such, Northeast Asia is better placed to acquire unmanned systems, and by extrapolation LAWS, through access to foreign sources or domestic development, compared to other sub-regions.

In Southeast Asia, Indonesia, Malaysia, the Philippines, Singapore, Thailand, and Vietnam not only diversify their sources of semi-autonomous military technologies, but have also attempted to build self-reliance through indigenous development programmes. Of these, Singapore is arguably the most successful, having produced the Venus USV and also having developed an innovative unmanned hybrid vehicle (UHV), touted "air-phibious" drone, capable of aerial and undersea operations.[22] Vietnam is developing a HALE drone, reportedly with Belarussian assistance.[23]

In South Asia, India and Pakistan are the only countries with major semi-autonomous systems programmes, drawing lessons from the use of foreign drones and, especially for Pakistan, most possibly with Chinese assistance. The close Sino-Pakistani military relationship, short of an alliance, does appear to give Islamabad much leeway in securing long-

22 C.P. Cavas, Dawn of the Air-Phibious Drones? Flying Fish UAV Swims and Flies, Defense News, 18 February 2016.

23 R.D Fisher, Jr., New Vietnamese HS-6L HALE UAV Likely Aided by Belarus, Jane's Defence Weekly, 23 December 2015.

term support from Beijing to sustain its programme. India's ambitious semi-autonomous systems programme furthermore includes an indigenous unmanned combat aerial vehicle (UCAV) dubbed the Autonomous Unmanned Research Aircraft (AURA), slated to be operational by 2023 contingent on funding availability.[24]

Operational deployment of semi-autonomous systems in the maritime domain

Operational deployment of semi-autonomous systems in the maritime domain may offer a helpful lens with which to view how LAWS will possibly feature in the Asia-Pacific littorals' future use of military force.

The most explosive of such maritime incidents involving semi-autonomous systems took place in Northeast Asia. In September of 2013, an unidentified drone was detected flying close to the disputed Senkaku/Diaoyu Islands, and well within the Japanese air defence identification zone though it kept clear of Japanese territorial airspace. Upon being intercepted by Japanese fighter jets, the UAV made a u-turn and headed in the direction of China.[25] This incident was the debut appearance of semi-autonomous systems in the Sino-Japanese dispute since tensions flared up in late 2012 over Tokyo's nationalisation of the isles. Following this incident, Japanese media reported that Tokyo has authorised the Self Defense Forces to shoot

24 M. Pubby, Government Set to Clear Rs 3,000 Crore Plan to Develop Engine for India's First UCAV, The Economic Times, 15 November 2015.

25 An unnamed Japanese defence official told the press that the drone's nationality was unclear, but added that it flew in from the northwest and was last observed flying towards that direction. However, a picture released by the Japanese Defense Ministry appeared to show a Chinese-made UAV. The Chinese National Defence Ministry effectively admitted in a statement on the night of the incident that the drone belonged to its military, but it later backtracked and claimed that Chinese military aircraft had never infringed on another country's airspace; see "Plane Thought to Be Drone Flies Near Senkaku Islands", Kyodo News, 9 September 2013; "Japan Scrambles Jets for Drone Near Disputed islands", Agence France Presse, 9 September 2013; "Japan Spots Unidentified Drone Near Senkakus", Jiji Press English News Service, 9 September 2013; Foreign Ministry Spokesperson Hong Lei's Regular Press Conference on 9 September 2013, Ministry of Foreign Affairs of the People's Republic of China; "Japan Calls on China to Curb Unmanned Flights Near Senkakus", Kyodo News, 10 September 2013; "Japan Coast Guard Officers May Have Spotted China Drones Before", Kyodo News, 10 September 2013; "Japanese Spot UAV Operating over Disputed Islands", Forecast International Defense Intelligence Newsletters, 10 September 2013; "China Defends Military Training in W. Pacific", Xinhua News Agency, 26 September 2013.

down foreign drones that intrude into Japanese airspace and fail to heed warnings to leave,[26] prompting the Chinese defence authorities to issue a warning.[27] Since this one event, however, no more drone intrusions over the East China Sea have been reported.

The next incident took place the following year in the Korean Peninsula. In March 2014, shortly after the North Korean military had staged a three-hour artillery live-firing drill along the Northern Limit Line – a de-facto maritime boundary enforced by Seoul but contested by Pyongyang, in the Yellow Sea where numerous inter-Korean naval skirmishes had taken place – the South Koreans recovered an unidentified UAV which had crashed into the nearby island of Baengnyeong. It was believed to be a North Korean drone. This sparked off a round of fiery inter-Korean accusations, heightening tensions on the Peninsula.[28] More than five months later, the wreckage of a suspected North Korean UAV was recovered on the same island.[29] In January 2016, South Korean troops fired warning shots following the intrusion of a North Korean drone across the Military Demarcation Line.[30] This episode highlights the potential risk of escalation into armed confrontation as a result of drone use.

Southeast Asia's situation has been much calmer. In September 2012, following the warning by Philippine defence authorities that Chinese drones may be shot if they enter the adjacent waters of the disputed Spratly Islands and Scarborough Shoal – which had been occupied by Chinese forces in April that year following a standoff with the Filipinos – China's Ministry of Defence warned against any military provocation and defended its UAV surveillance flights.[31] However, in July-August 2014, the Philippine garrison on the Second Thomas Shoal, also contested by China, monitored at least three overhead passes of a kind of UAV, which coincided with the increased presence of Chinese vessels off the shoal. However, Manila's threat of

26 "Japan to Down Intruding Foreign Drones if Warnings Ignored", Kyodo News, 20 October 2013.

27 "China to Strike Back if Japan Shoots Down Unmanned Aerial Vehicle – Official", BBC Monitoring Asia Pacific, 26 October 2013.

28 K. Eun-Jung, "S. Korea Probing Unknown Drone Found on Western Border Island", Yonhap English News, 1 April 2014; "Drone Crashed in Baengnyeong Took Photos of Border Islands", KBS World News – English Edition, 3 April 2014; "South Refutes N. Korea's Denial of Drone Incursion", Korea Times, 14 April 2014; K. Eun-Jung, "S. Korea Confirms Three Drones Were from N. Korea", Yonhap English News, 8 May 2014.

29 "Wreckage of Suspected NK Drone Found Near Border Island: Military", Yonhap English News, 15 September 2014.

30 K. Hyung-Jin, "S. Korea Fires Warning Shots After North Korean Drone Seen", Associated Press Newswires, 13 January 2016.

31 This happened after China's State Oceanic Administration announced plans to use UAVs to strengthen Beijing's maritime surveillance efforts over the disputed waters; see "China to Deploy Drones for Marine Surveillance", Xinhua News Agency, 29 August 2012; "DM Defends China's South China Sea Drones", Xinhua News Agency, 27 September 2012.

shooting at the drones was not followed through.[32] An interesting episode took place when it was revealed that in 2012, a Chinese fisherman had recovered a foreign UUV of unknown nationality, off Hainan Island, around the time tensions started to flare up.[33]

Thus far, no maritime incidents involving semi-autonomous systems have happened in South Asia. However, between India and Pakistan there has been a recurring history of drone intrusions across the land border since the early-2000s. It is therefore not presumptuous that Indo-Pakistani 'unmanned war' may extend into the maritime domain in the foreseeable future, considering that both countries have been keen to expand their drone arsenals. No such destabilising use of semi-autonomous systems has happened in Oceania.

Conclusions: grounds for optimism?

Taking the above four influencing factors together, it becomes clearer that the Asia-Pacific presents an uneven picture as regards the future prospects of LAWS utilised in the maritime domain. Northeast Asia appears to be the more likely sub-region to witness the proliferation and actual use of LAWS. This is less so the case for the other sub-regions. The longstanding absence of structural arms control to limit the type and quantity of naval armaments only means that current spate of military build-ups in the region will continue unchecked, thereby heralding the possibility of LAWS being incorporated in the arsenals.[34] Even operational arms control, which seeks to build transparency and confidence among governments and their militaries by placing limitations on how and where military capabilities can be deployed, has been found sorely lacking success and progress in the Asia-Pacific.

That being said, perhaps there is a silver lining. Notwithstanding the military build-ups which may be extrapolated to imply the possible proliferation of LAWS in varying, uneven extents throughout the Asia-Pacific littorals, it may be geography that mitigates the proliferation and actual employment of LAWS. As Figure 2 illustrates, the Asia-Pacific maritime

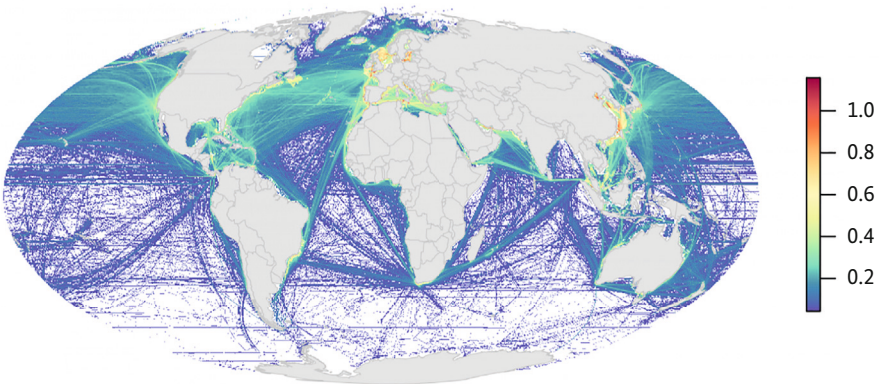
32 J. Laude, "Chinese Drones Fly Over Phl ship in Ayungin", The Philippine Star, 21 August 2014.

33 Q. Xie, "Chinese Fisherman Finds Metre-Long ‚Torpedo' in Controversy Plagued South China Sea… Which Turns out to Be an Under-Water Spy Device", Mail Online, 24 August 2015.

34 This is if one considers that many Asia-Pacific countries are State Parties to the Convention on Certain Conventional Weapons (CCW) – Australia, Bangladesh, Cambodia, China, India, Laos, New Zealand, Pakistan, the Philippines, South Korea, the Russian Federation, and Sri Lanka – but almost as many which are either just State Signatory (Vietnam), or not parties to the convention: Brunei Darussalam, Indonesia, Malaysia, Myanmar, North Korea, Singapore, Thailand, and Timor-Leste. States parties and signatories to The Convention on Certain Conventional Weapons, The United Nations Office at Geneva, http://www.unog.ch/80256EE600585943/(httpPages)/3CE7CFC0AA4A7548C12571C00039CB0C?OpenDocument.

geography, tied so intricately close to the economic interdependence among the regional countries and with the rest of the international community, invariably means that the potential risks of LAWS – both geopolitical destabilisation and dangers of collateral casualties and damage – can have an adverse, reverberating effect on global peace and stability.

Figure 2: Shipping Density Worldwide



Source: National Center for Ecological Analysis and Synthesis, *A Global Map of Human Impacts to Marine Ecosystems*: https://www.nceas.ucsb.edu/globalmarine/impactbyactivity.

The very geographical nature of the Asia-Pacific littorals – characterised by semi-enclosed and open waters plied by dense civilian sea traffic carrying the bulk of the world's trade shuttling between the Indian and Pacific Oceans – does serve a deterrent against the actual commitment of LAWS into real military operations, even if one assumes that such a technology may still proliferate among the regional militaries. Some of the Asia-Pacific governments at least are cognisant of the potential dangers contemporary naval armaments pose to civilian shipping in a cluttered maritime environment.[35]

35 One example was Singapore, which once mulled the possibility of acquiring a supersonic anti-ship missile to equip its navy frigates, perhaps inspired by her neighbours back in the early-2000s. However, there were apparently concerns about its feasibility given the heavy sea traffic and therefore cluttered surface background, which could pose dangers for such a weapon to hit an innocent ship instead of a legitimate target. Since then, nothing was heard about Singapore's intent to acquire such a capability, even after the Indonesians successfully tested a Russian-supplied supersonic anti-ship missile in April 2011; see Keynote Address by P. Ho, Permanent Secretary (Defence), at the Naval Platform Technology Seminar 2003, held at Singapore Exposition Conference Hall, Ministry of Defence, Singapore, 11 November 2003.

If Asia-Pacific state-centric scenarios concerning LAWS in the maritime domain are judged to be less likely, what about non-state actors? While LAWS may be costly, terrorist groups may still find it attractive to acquire just a small handful, which may not be as technically sophisticated as those designed for militaries. The non-state LAWS threat may plausibly emanate from aerial and underwater use. Nonetheless, many Asia-Pacific governments have adopted regulations governing the acquisition and use of civilian UAVs – which have found increasing attention in the commercial world. Such moves are designed to deal with the potential terrorist UAV use.[36]

It is likely that Asia-Pacific governments will remain keen on implementing measures to prevent or mitigate the proliferation and use of LAWS by non-state actors. The historical difficulties encountered in arms control, and instances of failures in implementation, mean that the interstate divergences regarding the types and functions of armaments, especially those that operate in the maritime domain, look set to persist into the future. This will foreseeably extend into the realm of LAWS.

The onus thus falls on states exercising unilateral restraint in the acquisition and employment of such a technology, considering the stakes involved now and in the future, while at the same time preserving and enhancing the efficacy of regional security mechanisms that seek to maintain peace and stability in the Asia-Pacific.

36 For example, in March 2016, Japan enacted the Civil Aeronautics Law, banning drones from flying over important facilities such as the Prime Minister's Office and giving police the power to destroy the drones if necessary. The move came after a small drone was found carrying a container filled with radioactive soil on the roof of the building housing the PMO in April 2015; see "Japan enacts tough new law to regulate drones", The Japan Times, 17 March 2016.

International and Regional Threats Posed by Lethal Autonomous Weapons Systems (LAWS): A Russian Perspective

Vadim Kozyulin*

Introduction

The development of autonomous technologies might cause some risks that should be identified and neutralised. The emergence of autonomous weapons systems (AWS) will increase tension, and have especially dangerous impact in regions with protracted territorial disputes or political frictions.

Threat of Occasional Incidents

Completely autonomous weapons such as the automated sentry gun, unmanned surface vehicles, or drone aircraft are beginning to appear in military arsenals. One day the AWS might be developed to the level that could pose a threat of occasional incidents in the air, sea, or deep waters.

There were cases in the past when lethal weapons were confused and malfunctioned due to certain unpredicted circumstances. For instance, it was reported that eight new F-22 fighter jets experienced total computer failure when crossing the international date line.[1]

Some specialists are concerned about the ability of AWS to abide by maritime traffic conventions, air traffic rules, and legal restrictions.[2] Stealthy AWS could be of a particular threat while dwelling through inhabited areas, sea, and air communication routes. Autonomous weapons systems will inevitably lack the flexibility that humans have to adapt to

* PIR Center for Policy Studies, Moscow.

1 B. Hill, Lockheed's F-22 Raptor Gets Zapped by International Date Line, Daily Tech, 26 February 2007, http://www.dailytech.com/Lockheeds+F22+Raptor+Gets+Zapped+by+International+Date+Line/article6225.htm.

2 S. Cheney-Peters, Insights into Unmanned ASW, CIMSEC, 24 August 2012, http://cimsec.org/insights-into-unmanned-asw-actuv/2320.

novel circumstances. One day, some unmanned surface or subsurface vehicle could cause an accident, and it is important to predict the legal and operational fallout in order to prevent regional or international implications.

Cyber security failures

Incidents might also occur due to loss of communication, jamming, control interception, or, in other words, because of communication security or cyber security failures. Such weapons could be uncontrollable in real-world environments where they are subject to design failure as well as hacking, spoofing, and manipulation by adversaries.[3]

Armed forces and security agencies in many countries dispose a wide spectrum of electronic warfare equipment that is able to conduct such activities as electro-optical, infrared, and radio frequency countermeasures; EM compatibility and deception; communications jamming, radar jamming, and anti-jamming; electronic masking, probing, reconnaissance, and intelligence; electronics security; EW reprogramming; emission control; spectrum management; and wartime reserve modes.[4]

It is therefore crucial to discuss the improvement of control systems of military AWS in order to prevent incidents. The issue of control systems reliability is closely connected to the issue of cybersecurity as it includes encryption, security of communication and data, development of standards that ensure reliability of communication, and data processing facilities.

Terrorist drone attacks

There is some evidence that terror groups examine possibilities for using unmanned drone aircraft in order to attack public gatherings. For decades, rebels and other non-state armed groups have used remotely operated drones for reconnaissance purposes. For instance, there have been 14 reports about instances of drones flying over nuclear power plants in France.

3 J. Markoff, Report Cites Dangers of Autonomous Weapons, The New York Times, 28 February 2016, http://www.nytimes.com/2016/02/29/technology/report-cites-dangers-of-autonomous-weapons.html?_r=0.

4 U.S. Joint Chiefs of Staff, Electronic Warfare, Joint Publication 3-13.1, 25 January 2007, https://fas.org/irp/doddir/dod/jp3-13-1.pdf.

In 2014, the NYPD, the largest police force in the United States, became was increasingly concerned about a potential terror attack from the air by a drone armed with a deadly weapon. In one particular incident, an NYPD Aviation Unit helicopter almost collided with a drone that was about 800 feet above ground.[5]

Another potential issue in this context is the possibility to disseminate incendiary pamphlets by means of a remotely controlled aircraft, which could cause clashes among the civilian population. One day, furthermore, terrorists might be sophisticated enough to carry out attacks with chemical or biological weapons. Terrorists could also turn remotely-piloted aircraft into flying bombs by hooking them up to improvised explosive devices.[6]

In the summer of 2013, German law-enforcement personnel raided Islamic militants and right-wing extremists believed to be plotting drone attacks. In the course of the operation, police recovered bomb-making materials and a drone from the right-wing extremists, who were allegedly planning to use the device to bomb a German summer camp.

A spokesperson for the U.S. Federal National Counterterrorism Center has said that "[e]fforts by terrorists to use drone technology are obviously a concern (...). Our focus remains on identifying these threats and supporting those agencies who are responsible for countering them".[7]

Global cities are increasingly the targets of terrorist attacks. Meanwhile, numerous regional conflict areas are particularly vulnerable to provocations that can spark religious, ethnic, or international conflict.

5 NYPD Scanning the Sky for New Terrorism Threat, CBS News, 29 October 2014, http://www.cbsnews.com/news/drone-terrorism-threat-is-serious-concern-for-nypd/.

6 B. Farmer, Terrorist 'Could Use Drones for Chemical and Biological Attacks, The Telegraph, 22 October 2014, http://www.telegraph.co.uk/news/uknews/terrorism-in-the-uk/11177388/Terrorist-could-use-drones-for-chemical-and-biological-attacks.html.

7 J. Nicas, Use of Drones by Criminals and Terrorists on the Rise, The Wall Street Journal, 29 January 2015, http://www.matthewaid.com/post/109479490261/use-of-drones-by-criminals-and-terrorists-on-the.

UAVs Sow Seeds of Terrorism

Deployment of autonomous weapons systems in local regional conflicts will allow operators, commanders, and politicians who sanction the application of these weapons to remain out of combat and beyond the reach of the adversary. Deprived of the opportunity to counterattack in open battle, the adversary will have to resort to other available means. Terrorism might be the most common reaction as probably the only possible asymmetric response to the attacks by the high-tech opponent. The community subjected to LAWS attacks will inevitably cultivate hatred of those inaccessible and untouchable people who use robots to attack while remaining in a safe place.

The deployment of combat drones by the United States in Pakistan and Yemen in the past decade might give us an illustration of a similar situation. Leaked military documents reveal that the vast majority of people killed by the unmanned aerial vehicles were not the intended targets, with approximately 13 percent of deaths being the intended targets, 81 percent being other militants, and 6 percent being civilians. An estimated of 423 to 965 civilians have been killed in the so-called 'drone war', including 172 to 207 children, with 2,497 to 3,999 persons killed in total.[8]

Human Rights Watch and Amnesty International have documented their findings on specific drone strikes in Yemen and Pakistan. The report determined that in six air strikes in Yemen, 57 of the 82 casualties were civilians, including a pregnant mother and three children. In Pakistan, more than 30 civilians died in four similar strikes. A UN human rights investigator's report concludes that drone strikes, no matter how precise they are claimed to be, can still result in significant collateral damage.[9]

Many drone technicians believe that drone attacks fuel hatred in the victims' families. And as they simply do not have anywhere else to turn, chances are they turn to radical organisations. Investment in economic development in Afghanistan to provide opportunities for working adults and youths could be a more effective way to defeat terror. Investment in gratitude and success would be the right approach instead of breeding a generation full of fear and hatred.[10]

8 See The Bureau of Investigative Journalism, Get the Data: Drone Wars, https://www.thebureauinvestigates.com/category/projects/drones/drones-graphs/.

9 D. Donovan, The Drone Disaster, U.S. News & World Report, 13 February 2014, http://www.usnews.com/opinion/blogs/world-report/2014/02/13/american-drones-are-creating-terrorists-in-africa-and-asia.

10 Ibid.

Increasing potential for surprise offences

While many countries conduct intensive military research and development, some new technologies was the result of a significant and targeted military funding of science. Extensive funding of the military technology permits to outmarch competitors and to gain crucial dominance on the battlefield, or strong positions for a surprise attack. Here are some most advanced technologies which can change the nature of modern war. Altogether, they shape an autonomous military mechanism which increases potential for a surprise offensive.

Anti-Submarine Unmanned Vessel:[11] an unmanned vessel optimised to robustly track quiet diesel electric submarines for months at a time, spanning thousands of kilometres of ocean with minimal human input. The vessel will be small and cheap (target cost goal of $20 million apiece), yet robust enough to operate for 80 days and 6,200 kilometres without human maintainers or refuelling.[12]

Submersible Combat Sensing Gliders might become a permanent underwater partner of the Anti-Submarine Unmanned Vessels.[13] The Glider is a long endurance UUV that is propelled by changes in buoyancy along with its wings and tail-fin steering. The gliders collect oceanographic data and can operate up to 30 days and even much more at a time on a lithium battery.[14] It has obvious uses for submarine hunting and hiding, given the effect of temperature layers on sonar propagation. In 2009, a Slocum Glider managed to cross the Atlantic in 223 days.[15]

Hypersonic vehicles capable of delivering nuclear warheads at more than 10 times the speed of sound, or 12,231.01kp. The missile defences can only counter ballistic missiles and warheads that have predictable trajectories, whereas hypersonic vehicles are capable of manoeuvring during flight, and can reach any point of the globe within two hours while travelling at the edge of space, which makes them extremely difficult to shoot down.

11 See http://www.darpa.mil/program/anti-submarine-warfare-continuous-trail-unmanned-vessel.
12 Cheney-Peters (n 2).
13 See http://www.naval-technology.com/contractors/sonar/teledyne-webb-research/pressteledyne-reaches-third-milestone.html.
14 See http://www.navaldrones.com/LBS-Glider.html.
15 See http://www.defenseindustrydaily.com/underwater-gliders-for-the-us-navy-06990/.

Refuelling UAVs, autonomous unmanned aircraft capable of being refuelled from a tanker while airborne. They have the potential to conduct bombing missions over long distances.[16] The refuelling drones are the first step towards an unmanned, stealth strike aircraft that could conduct missions over much longer distances than fighter jets — allowing aircraft carriers to operate outside of the range of the enemy missiles.

Military Micro-Drones: a swarm of micro-drones that can be dropped from an airplane to glide down to the earth's surface, directed by an on-board GPS which can guide the drones to its target destination from a 17 kilometre descent.[17] The mini-drones can accommodate different payloads, and communicate both with other drones and the manned aircraft acting as their mothership. An airplane going in against an integrated air defence system will drop 30 drones that subsequently form a network and act as sensors and/or decoys.[18] Upon accomplishing their mission, the drones will return to the mothership and will be ready for the next mission within 24 hours.

Military Swarming or *Centaur Warfighting*: military swarming involves human-machine teaming and 'swarming' operations by unmanned drones as well as the use of a decentralised force against an opponent, in a manner that emphasises mobility, communication, unit autonomy, and coordination or synchronisation.[19] This battlefield tactic is designed to overwhelm or saturate the defences of the principal target or objective. There is now active research in consciously examining military doctrines that draw ideas from swarming.

16 G. Dyer, US Military: Robot Wars, Financial Times, 7 February 2016, https://www.ft.com/content/849666f6-cbf2-11e5-a8ef-ea66e967dd44.

17 M. Santos, The Coming Age of the Military Micro-Drone, Futurism, 2 February 2016, http://futurism.com/the-coming-age-of-the-micro-drone/.

18 See http://www.americanmilitaryforum.com/forums/threads/perdix-mini-air-launched-swarming-uav.579/.

19 S.J.A. Edwards, Swarming on the Battlefield: Past, Present, and Future, RAND Corporation, 2000, http://www.rand.org/pubs/monograph_reports/MR1100.html.

War dehumanisation by diminishing human engagement

Engagement of remotely controlled and automated weapons reduces human engagement in combat and diminishes the psychological impact of the casualties inflicted by the operators. The factor of human participation at war gradually vanishes as military robots replace human soldiers on the battlefield. Such dehumanisation of war, i.e. the reduction of human participation in military conduct, appears preferable because it allows for the reduction of manpower, salaries, healthcare, and pension bills.

In some armed forces, robots may replace one fourth of combat soldiers by 2030.[20] While not explicitly stated, a major motivation behind replacing humans with robots is that humans are expensive. Training, feeding, and supplying them while at war is pricey, and after the soldiers leave the service, there is a lifetime of medical care to cover.[21] Further reducing human involvement in military service could result in distancing of the human society from sufferings of those being defeated by the high tech army.

Provoking ignorance of international law

Modern AWS open the door for non-contact warfare that allows for the saving of troops and inflicting smaller damage to friendly forces while fighting an archaic enemy. Deployment of drones, smart weapons, electronic warfare, and other advanced means of war lead to quick advance and easy victories. In general, advanced weapons provide politicians with a reason to ignore the international law for the sake of fast achievement of their goals.[22]

There have been some instances of military conduct based on high-tech dominance strategies without UN approval:

20 K.D. Atherton, Robots May Replace One-Fourth of U.S. Combat Soldiers by 2030, Says General, Popular Science, 22 January 2014, http://www.popsci.com/article/technology/robots-may-replace-one-fourth-us-combat-soldiers-2030-says-general.

21 Ibid.

22 International Committee of the Red Cross, The Use of Armed Drones Must Comply with Laws, Interview with Peter Maurer, 10 May 2013, https://www.icrc.org/eng/resources/documents/interview/2013/05-10-drone-weapons-ihl.htm.

- *Yugoslavia, 1995*: NATO military operation without UN approval; the first large-scale military operation in the history of NATO violated the principles of international law, as the UN Security Council had not adopted a resolution that permitted the use of military force by NATO countries.
- *Afghanistan and Sudan, 1998*: a military strike unilaterally carried out by the U.S. In 1998, large-scale terrorist attacks against the U.S. embassies in Kenya and Tanzania took place. According to U.S. intelligence officials, the attacks had been organised by the then little-known terrorist group al-Qaida. In response to these attacks, U.S. President Bill Clinton ordered air strikes with cruise missiles against al-Qaida's camps in Afghanistan and a pharmaceutical factory in Sudan.
- *Yugoslavia, 1999*: NATO intervention without UN sanction. The war in Kosovo, which had begun in 1996, became the pretext for the armed intervention by the U.S. and NATO. Like the air strikes against the Bosnian Serbs in 1995, the operation against Serbia was positioned by Washington as a 'humanitarian intervention'. As part of this 'humanitarian intervention', NATO aircraft attacked military infrastructure as well as Serbian cities, bridges, and industrial enterprises over the course of almost two and a half months. Belgrade and other major cities were subjected to missiles and airstrikes.
- *Iraq, 2003*: the U.S. and its allies intervened in Iraq without UN approval. Washington used forged evidence and false intelligence in order to convince the world that Iraq was actively developing weapons of mass destruction. However, representatives of Russia, France, and China made it clear that they would veto any draft resolution for the use of force against Iraq. In March 2003, the U.S. and allied states launched operation 'Iraqi Freedom'.
- *Libya, 2011*: the NATO intervention was authorised by the UN Security Council. In February 2011, riots had begun in Libya, which escalated into full-scale armed conflict between opposition groups and government forces led by Muammar Gaddafi. In late February 2011, the UN Security Council adopted a resolution that imposed sanctions against the Libyan government. In March 2011, another resolution installed a no-flight zone over the territory of Libya. Subsequently, NATO countries started bombing positions of government forces and military infrastructure. Though the civil war in Libya officially ended in October 2011, armed clashes between paramilitary groups and various militias continue to this day.

Considerable military superiority and the ability to deliver damage to the enemy armed forces without close encounters permit political leaders to use their military potential as a universal leverage for solving international problems, without the approval of the United Nations, and more often than not under violation of the rules of the *jus ad bellum*.

Proliferation of LAWS

Transfers of sensitive technologies used for production of certain AWS, such as attack drones, may considerably destabilise the situation in conflict zones and disputed areas. The proliferation of UAV technologies can serve as a destabilising factor, raising tension in numerous local territorial conflicts.

In this context, it is worth noting that global expenditure on military UAVs will achieve revenues of $7,447 million in 2016.[23] The Teal Group predicts that the total worldwide UAV market will more than double over the next decade, from current expenditures of $5.2 billion annually to $11.6 billion, totalling just over $89 billion over the next 10 years. Uncontrolled sales of UAVs fuel conflicts in many disputed areas of the world. Transfer of LAWS is yet another phenomenon that rapidly increases regional and international risks.

Temptation of global superiority

Rapid technological changes give a chance for a global military dominance which persuades major powers to rely on military force and global technological superiority with regard to finding solutions for security problems and for maintaining deterrence. Such rhetoric alarms the other powers' leaderships and whips up their military spending and defence preparations.

Breaking existing military balances

Some military planners believe that smart, robot weapons can help to restore deterrence that has eroded as of late.[24] The underlying objective of the new strategy is to find weapons and technologies that ensure that friendly forces can evade the layered hostile missile defence, to defend bases against attack from precision-guided missiles, and to be able to operate carrier fleets at a much greater distance from an enemy. In the meantime, nuclear

23 Military Unmanned Aerial Vehicle (UAV) Market Report 2016-2026, PR Newswire, 11 April 2016, http://www.prnewswire.com/news-releases/military-unmanned-aerial-vehicle-uav-market-report-2016-2026-575245631.html.

24 D. Ignatius, The Exotic New Weapons the Pentagon Wants to Deter Russia and China, The Washington Post, 23 February 2016, https://www.washingtonpost.com/opinions/the-exotic-new-weapons-the-pentagon-wants-to-deter-russia-and-china/2016/02/23/b2621602-da7a-11e5-925f-1d10062cc82d_story.html?utm_term=.ace156d65337.

deterrence between great powers is even further marginalised as a factor in international security.[25]Some military strategists believe that AWS can help restore deterrence and replace nuclear weapons and precision-guided conventional weapons. However, wide deployment of AWS at tactical and strategic levels could destroy military balances and increase the risks of conflicts by provoking a pre-emptive strike.

Raising regional and international tension

The deployment of the controversial U.S. missile defence system in combination with the development of AWS and hypersonic weapons provokes the increase of global military expenses. The present escalation of official statements opens the door for mutual accusations and speculations which gradually make the public opinion believe that potential military conflicts between nuclear powers are inevitable and might happen soon.

Provoking conflicts

The disturbed international political environment could be a rather responsive space for provocations initiated by the use of autonomous weapons, or accidents caused by technical failures or operator's mistakes.

Wide application of air and underwater drones might lead to the erosion of escalation threshold, as unmanned vehicles are more affordable than other military aircraft. And with no human pilot at risk, drones could make it easier to decide to go to war. "The proliferation of this technology will mark a major shift in the way wars are waged", in the words of Daryl Kimball, the executive director of the Arms Control Association. "We're talking about very sophisticated war machines here. We need to be very careful about who gets this technology. It could come back to hurt us."[26]

25 C.M. Leah, There Will Be More Cold Wars, Business Insider, 7 April 2016, http://www.businessinsider.com/there-will-be-more-cold-wars-2016-4?IR=T.

26 J. Wolverton, II, U.S. Drone Manufacturers Contribute Millions to Congressional Campaigns, The New American, 14 July 2012, http://www.thenewamerican.com/usnews/foreign-policy/item/12078-us-drone-manufacturers-contribute-millions-to-congressional-campaigns.

Militarisation of civil tech sector

The new autonomous weapons age dictates a new approach to technological innovations and a close alliance with the most advanced high-tech corporations. It makes militaries cultivate and facilitate a lasting relationship with new innovators, and those who do not always work together with the U.S. Department of Defense, to help expand its so-called 'innovative ecosystem of ideas'. In order to get their hands on the latest software technology to manipulate and take advantage of large volumes of data, military strategists invest in start-ups, award classified contracts, and recruit technology experts. In this sense, the world is witnessing the rise of something we might call the Military-Information Complex.[27]

The lack of transparency

The sphere of AWS designing and production is very sensitive and secretive in every country. The production of AWS engages design companies which consume millions of dollars for the creation of navigation and mapping technologies, software, optics, etc. Some spheres, specifically the areas of weapons, artificial intelligence, night-vision goggles, micro-devices, and communications, belong to the most preserved state secrets. The globally kept information taboo on the most advanced technological achievements in the military sphere fuels suspicions and tension in the relations between states.

New arms race

In an effort to match the huge defence budgets of rival states, the other global military powers have to increase their defence spending in order to gain the most advanced military technologies.[28] There are signs that the further development of offensive AWS could lead to a 'new arms race'.

Unable to compete with the United States as regards the development of the Anti-Ballistic Missile Defense or autonomous weapons, the Russian Federation makes an attempt to find a so-called 'asymmetric response' to the potential of a threat by the U.S., for instance

27 B. Merchant, The Military-Information Complex Is Growing in Silicon Valley, Motherboard, 20 June 2013, http://motherboard.vice.com/blog/the-military-information-complex-is-rising-in-silicon-valley.

28 H.S. Friedman, 5 Countries with the Highest Military Expenditure, The Huffington Post, 29 November 2011, http://www.huffingtonpost.com/howard-steven-friedman/military-spending-united-states_b_1118851.html.

by developing hypersonic missiles or designing the long-range, nuclear-armed torpedo 'Status-6', which might be deployed in autonomous mode.[29] While experts dispute whether this torpedo is real or not, one may assume that this phenomenon signals an escalation of the arms race which might unleash the most deadly and ruinous means of destructions.

Threat of sliding into World War III

The development of autonomous robotic systems and other sophisticated weaponry evokes the events that took place a century ago. In the late 19th century, the European general headquarters were inspired by new technological innovations. The newly constructed railroads were the most promising military means as they allowed to quickly deliver a large number of troops over long distances in a short time. The second innovation that added adrenaline to military strategists were mobilisation plans, invented in Prussia and adopted by other powers. The telegraph was the third revolutionary invention, which linked military units and institutions and let the mobilisation plan go through much faster. The fourth innovation – the general conscription – allowed for the keeping of large, trained military reserves.

These inventions radically increased military capabilities of states powers, and eventually led to disaster. It turned out that after the mechanism of build-up triggered, it was almost impossible to make it reverse. On 16 July 1914, a week after Germany had begun covert mobilisation, Russia announced a partial mobilisation as well. On the same day, Germany presented Russia with an ultimatum: stop conscription into the army, or Germany will declare war on Russia. Russian generals convinced the Emperor that the cessation of mobilisation will leave Russia naked against the German troops coming to the border. The ultimatum was rejected.

There is no doubt that combat robots will significantly increase the military potential of powers, and AWS mass adoption could be a threat to the military balance in different regions, ultimately increasing risks of conflict. It is crucial for mankind to assess and neutralise this threat in order to let artificial intelligence and autonomous robots become a technology for the benefit of a safer world, and not synonymous with an unknown and irreversible danger.

29 S. Pifer, Russia's Perhaps-Not-Real Super Torpedo, Brookings Institution, 18 November 2015, https://www.brookings.edu/blog/order-from-chaos/2015/11/18/russias-perhaps-not-real-super-torpedo/.

Autonomous Weapons Systems and the Obligation to Exercise Discretion

Eliav Lieblich*

Introduction

This chapter argues that a key problem posed by AWS[1] is that they constitute a use of administrative powers against individuals without the exercise of proper discretion. AWS are based on pre-programmed algorithms, and therefore – as long as they are incapable of *human*-like metacognition – when they are deployed, administrative discretion is bound. Operating on the basis of bound discretion is *per se* arbitrary and contradicts basic notions of administrative law, notions that, as argued here, complement modern standards of international humanitarian and human rights law. This realisation explains better some of the concerns relating to AWS, which are usually expressed in circular arguments and counter-arguments between consequentialist and deontological approaches.

That machines should not be making 'decisions' to use lethal force during armed conflict is a common intuition. However, the current discussion as to just why this is the case is unsatisfying. The ongoing discourse on AWS is essentially an open argument between consequentialists (instrumentalists) and deontologists. Consequentialists claim that if AWS could deliver good results, in terms of the interests protected by international humanitarian law (IHL), there is no reason to ban them. On the contrary, we should encourage the development and use of such weapons. Of course, proponents of this approach are optimistic about the ability of future technology to make such results possible. They also point out the deficiencies in human nature, such as fear, prejudice, propensity for mistake, and sadism that can be alleviated by autonomous systems.

* Assistant Professor, Radzyner Law School, Interdisciplinary Center (IDC), Herzliya. This presentation is based on E. Lieblich and E. Benvenisti, The Obligation to Exercise Discretion in Warfare: Why Autonomous Weapons Systems Are Unlawful, in N. Bhuta et al. (ed.), Autonomous Weapon Systems: Law, Ethics, Policy, Cambridge 2016, pp. 245-283; and E. Lieblich and E. Benvenisti, Autonomous Weapons Systems and the Problem of Bound Discretion, 38 Tel Aviv University Law Review (forthcoming 2016).

1 I use the term 'autonomous weapons system' (AWS) rather than 'lethal autonomous weapons system' (LAWS) since questions relating to autonomous use of force arise whether or not such force is necessarily lethal.

Those who object to AWS on instrumental grounds, conversely, argue that in the foreseeable future AWS will not be able to satisfy modern IHL's complex standards – such as distinction and proportionality – and therefore will generate more harm than good.[2] They furthermore point out that AWS will also eliminate the good traits of humanity, such as compassion and chivalry, from the battlefield. While these concerns seem convincing, they fail to lay down a principled objection to AWS since they can always be countered, at least analytically, by resort to optimistic hypotheticals regarding future technologies,[3] as well as to negative examples of human nature on the battlefield, which are unfortunately abound. Thus, a substantive discussion of AWS must transcend speculative claims regarding their ability to deliver end results, whether these are based on future technologies[4] or on mutually offsetting arguments about human nature.[5]

Deontologists claim that even if AWS could deliver good immediate outcomes, their use should still be prohibited, whether on ethical or legal grounds. The deontological objections focus on the nature of the computerised 'decision-maker' and the human dignity of potential victims.[6] However, deontologists, too, are placed in an awkward position when confronted with extreme hypotheticals. For instance, they have to admit that even if AWS would be better than humans in mitigating civilian harm in warfare, greater loss of life is preferable to lesser loss of life, only because a machine is involved in the process.[7] Furthermore, deontological approaches to AWS are lacking in that they argue from notions of dignity, justice, and due process,[8] but they do not tell us how and why these are relevant, as such, in situations of warfare.

2 Instrumentalist objections discuss additional problems, such as the difficulty of assigning *ex post* responsibility and lowering the 'price' of warfare which can result in diminishing the restraint on the use of force. In this chapter, however, I focus only on the primary issue of protection of individuals *in bello*.

3 See e.g. K. Anderson and M. Waxman, Law and Ethics for Robot Soldiers, Policy Review 2012, http://www.hoover.org/research/law-and-ethics-robot-soldiers; compare P. Asaro, On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making, 94 International Review of the Red Cross 687 (2012) 699.

4 Asaro (n 3), p. 699.

5 See e.g., N. Chomsky and M. Foucault, Human Nature: Justice vs. Power – The Chomsky-Foucault Debate, New York 2006.

6 See e.g. Mission Statement of the International Committee for Robot Arms Control, 2009, http://icrac.net/statements/.

7 See L. Alexander and M. Moore, Deontological Ethics, in E.N. Zalta (ed.), Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/ethics-deontological/.

8 Asaro (n 3), 700-701; this is because according to Asaro, the essence of due process is "the right to question the rules and appropriateness of their application in a given circumstance, and to make an appeal to informed human rationality and understanding", ibid, 700.

The discussion, thus, is caught in a loop of utilitarian arguments and deontological retorts, both not entirely satisfying. This presentation offers a middle-way approach to the question, based on an *administrative* perception of warfare. In particular, it serves to bridge the theoretical gap between warfare and administrative concepts of justice and due process, usually understood to be applicable during peace time.

War as governance: modern warfare as an exercise of administrative power

In order to properly discuss whether notions of justice and due process are relevant to the issue of AWS we have to first address the nature of modern warfare. The basic argument presented here is that under contemporary international law, war must be understood as a form of governance[9] – a fact that spawns administrative-legal obligations.

It must be conceded that traditionally, justice and due process were foreign to the idea of war. Classic sources of the laws of war, such as the 1863 Lieber Code, treat the citizens on the other side as part and parcel of the 'enemy' in the sense that they were expected, as such, to suffer the hardships of war.[10] War was thus seen as a violent struggle between collective entities, in which there was no room for individualisation.[11] Such collectivisation contradicts the idea of individual agency and responsibility, and is of course hardly an example of justice or due process as the terms are regularly understood. Under such an assumption, therefore, it was possible to dismiss any special obligations of justice and due process between a state and the enemy's civilians. If we adopt this view, then AWS should not be assessed in light of such standards, as deontologists suggest.

However, this perception of war was perhaps sustainable when most conflicts were between equally-capable states. Nowadays, most wars are asymmetric conflicts between states and non-state actors. In many cases, armed force is used by advanced militaries against such actors in failing states or territories. This would be the likely scenario in which AWS would be deployed. For various reasons detailed elsewhere – namely, the absence of accountable sovereigns – such conflicts result in a significant gap in the protection of

9 See E. Benvenisti and A. Cohen, War Is Governance: Explaining the Logic of the Laws of War from a Principal-Agent Perspective, 112 Michigan Law Review 1363 (2013); E. Lieblich, Show Us the Films: Transparency, National Security and Disclosure of Information Collected by Advanced Weapon Systems under International Law, 45 Israel Law Review 459 (2012) 483.

10 General Orders No. 100: Instructions for the Government of the Armies of the United States in the Field (Lieber Code), Art. 21.

11 See L. Oppenheim, International Law, para. 58.

civilians.[12] This gap must result in some 'diagonal' responsibilities of protection between the potentially affected individuals and the attacker, who is in a position to decide their fate and is residually capable of protecting them.[13]

This suggests that we must view modern warfare as a form of exercise of state power vis-à-vis individuals rather than a horizontal engagement between equal sovereigns. We can thus understand state action during armed conflict as the exercise of administrative, executive action. Once it is perceived this way, warfare should be subjected to widely accepted notions of administrative law that governs executive decision-making.[14] Importantly, such obligations can be triggered even before full territorial effective control by the attacker, both due to the broadening understanding of 'control' in contemporary international law[15] and the emerging general duty of sovereigns to take other-regarding considerations in their decision-making processes.[16]

Therefore, the fact that certain armed conflict cross national borders does not, in itself, negate administrative-like responsibilities between a state and civilians on the other side. Does this in itself mean that states must treat 'enemy' civilians as their own? A reasonable answer could be found in the principle of equal moral worth, which requires that a state cannot treat 'enemy' civilians beneath minimum acceptable standards, in a manner that it would never treat its own.[17] In the specific context of AWS, we may ask whether states would be willing to use them in situations where their own citizenry could be affected by 'decisions' made by such systems. It seems that some states already answered this question in the negative.[18] Indeed, if states limit computerised decisions with regard to their own

12 See E. Lieblich with O. Alterman, Transnational Asymmetric Armed Conflict under International Humaitarian Law: Key Contemporary Challenges, Institute for National Security Studies, Tel Aviv 2005, http://www.inss.org.il/uploadImages/systemFiles/Transnational%20Asymmetric_full%20text.pdf, pp. 18-19.

13 See generally E. Benvenisti, Rethinking the Divide Between Jus ad Bellum and Jus in Bello in Warfare Against Nonstate Actors, 34 Yale Journal of International Law 541 (2009).

14 Importantly, with the expansion in the understanding of the notion of 'control' in international law, such obligations might be triggered.

15 ECtHR, Al Skeini v. UK, Appl. no. 55721/07, Judgment of 7 July 2011, paras 131–40; Human Rights Committee, General Comment 31, Nature of the General Legal Obligation on States Parties to the Covenant, UN Doc. CCPR/C/21/Rev.1/Add.13, 2004, para. 10.

16 E. Benvenisti, Sovereigns as Trustees of Humanity: on the Accountability of States to Foreign Stakeholders, 107 American Journal of International Law 295 (2013).

17 See D. Luban, Risk Taking and Force Protection, Georgetown Public Law and Legal Theory Research Paper no. 11- 72, 2011, pp. 12, 46.

18 For instance, Art. 15 of Council Directive (EC) 95/46 enshrines the right of every person "not to be subject to a decision which produces legal effects concerning him (...) which is based solely on automated processing of data"; Council Directive (EU) 95/46 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 1995 Art. 15(1).

citizens, it is questionable whether they could subject others to such decisions. The effect of an administrative perception of warfare over the question of AWS is thus clear: AWS would be subject to additional, residual constraints, even if such weapons would perform reasonably, in terms of immediate results, under IHL.

Indeed, while IHL does not directly refer to principles of administrative law, some traits of administrative-legal thinking can be found even in positive law. For instance, the duty to take 'constant care' in the conduct of military operations[19] is reminiscent of the administrative law notion of the obligation to exercise discretion when making decisions, which is central to our argument as well. As we argue, the administrative approach should inform our understanding of the 'constant care' standard.

Similar ideas can also be derived from international human rights law, which requires that the limitation or (in some cases) deprivation of rights be subject to due process in the course of limiting those rights. The administrative law perception can thus additionally inform our understanding of what constitutes 'arbitrary' deprivation of life during armed conflict.

AWS and the binding of administrative discretion

As aforementioned, the administrative law perception allows us to understand the duty to take 'constant care' as a requirement to exercise continuous discretion during hostilities. This obligation requires the administrative authority to consider each decision, within the confines of its legal authority, in light of the specific goals of the authorising norm, as well as the rights and interests of those affected in the specific circumstances. Of course, this duty implies a prohibition of fettering one's discretion in advance, since binding one's discretion negates the ability to consider each case individually and make appropriate adjustments when needed.[20]

The justifications for the obligation not to fetter one's discretion are twofold. The first stem from notions of human dignity, which require the executive to give due respect to the individual by considering her specific case.[21] The second justification relates to the quality of decision-making, and assumes that in the long run, good administrative

19 Art. 57(1) Additional Protocol I (API).

20 See e.g. British Oxygen v Minister of Technology, [1971] AC 610, HL (UK). See generally J. Jowell et al., De Smith's Judicial Review, London 2014.

21 Benvenisti (n 16), p. 314.

decisions cannot be made, in a complex world, without making constant adjustments.[22] This is due to epistemological limitations, which constrain human ability to prejudge or foresee complicated situations.[23]

Deployment of AWS contravenes this duty. This is because during hostilities, the duty to exercise discretion requires the active, ongoing intention not to inflict harm on civilians.[24] It requires the commander (and subordinates) to exercise discretion both when planning the attack and, importantly, during the attack, up to the last moment before the trigger is being pulled.[25] In our context, in an out-of-the-loop scenario, the duty to exercise discretion 'in the last moment' would have to be performed by the AWS. As aforementioned, AWS cannot engage in the metacognition required for 'true' discretion. Their use thus reflects the stringent binding of executive discretion in advance – through the pre-programmed algorithms that govern their behaviour.

Binding discretion during warfare through AWS seriously contravenes both rationales of the duty to exercise discretion. First, it runs counter to the obligation to give due respect to the individual, since at hand are life or death decisions in which the potentially harmed individual is not considered at all but is 'factored' into the pre-determined processes. Second, war is an extremely complex environment, by nature requiring constant adjustments. In such an environment, any operation based on rigid, pre-determined decision-making patterns is unlikely to yield good, all-things-considered, long-term results.

Some proponents of AWS claim that human discretion is indeed exercised, by stressing that it is embedded into the system through the human discretion of the programmers. Discretion, then, is exercised, but on a different *temporal* level.[26] However, this argument is unconvincing since it precisely points out the problem we highlight: that AWS do not (and cannot) exercise discretion in real-time, as an administrative legal perception requires.

22 And this remains true even if acting without exercising discretion would bring good results in this or that specific case. The duty to exercise discretion ifs concerned with long run decision-making quality.

23 See e.g. T.J. Barth and E.F. Arnold, Artificial Intelligence and Administrative Discretion: Implications for Public Administration, 29 The American Review of Public Administration 332 (1999) 338, 348-349; see also H.L.A. Hart, Discretion, 127 Harvard Law Review 652 (2013) 661-664.

24 See M. Walzer, Coda: Can the Good Guys Win, 24 European Journal of International Law 433 (2013) 437.

25 Art. 57(2)(a)(i), 57(2)(b) API.

26 See e.g. M.N. Schmitt and J.S. Thurnher, "Out of the Loop": Autonomous Weapons Systems and the Law of Armed Conflict, 4 Harvard National Security Journal 232 (2013) 266.

Even if the discretion is exercised by the deploying commander,[27] this will not change our conclusion. This is because it is unlikely that the deploying commander would be able to predict how the AWS operates, a problem aggravated by the well documented 'computer bias', which causes humans to rely heavily on computer decision-making when available.[28] If, conversely, commanders *are* able to foresee the exact manner in which the AWS operates, this would imply that the weapon is not sophisticated enough to satisfy the complex standards of IHL to begin with. Now, if we suggest that the system can be significantly adjusted in real time – up to the last minute – then the system is possibly not autonomous at all. A further related argument is that an AWS can be programmed to freeze in complex situations it cannot resolve. However, ascertaining that a certain situation is complex is in itself a substantive decision that requires discretion.

Indeed, one can raise the question whether this analysis would hold in cases where 'friendly AWS' are deployed. For instance, let us assume that an AWS is charged with rescue operations in hazardous areas. Can such system be weaponised for 'self'-protection? Of course, it seems unreasonable that machines be allowed to kill just in order to preserve themselves, since this, in essence, is tantamount to recognising a right to kill in the defence of property. In principle, such a 'right' might make sense only as an extension of the right to life of the persons the AWS sets out to rescue. If the AWS is bound to rescue person X, and person Y attempts to destroy the AWS, person Y is in essence killing person X. The AWS could then be justified in killing person Y in order to save X, as an instance of defence of others. However, this might all be true in schematic theoretical cases, but in real life, significant discretion must be exercised in order to ascertain that this is indeed the situation. In essence, such cases merge with the (very) controversial question of whether AWS can be deployed in law enforcement operations and to perform the complex standards of use of force applicable in such contexts. In sum, the duty to take constant care and thereby exercise discretion results in the conclusion that AWS – as long as they do not possess the ability to exercise true discretion – cannot be allowed to make final targeting or killing decisions.

27 Ibid., 267.

28 Barth and Arnold (n 23), p. 348. It should be clarified that AWS can be based on rigid programming yet still be unpredictable.

What if AWS are deployed in circumstances where only combatants or direct participants in hostilities are targeted?

Until now, the discussion focused on the relations between the state and the adversary's civilians. However, does this analysis hold even in cases where AWS are deployed in 'traditional', symmetric battlefield scenarios, in which it is clear that no civilians are present? Similarly, what if they are deployed in uncluttered environments such as in deserts, the open sea, or in space? Indeed, some proponents of AWS claim that for a weapon to be banned, it needs to be shown that its deployment would be unlawful in *all* circumstances.[29] If, the argument goes, combatants can be targeted at all times merely by virtue of their status, then AWS might be lawful in uncluttered environments in which only combatants are present. If, however, there are some administrative-like obligations between the attacker and enemy troops, then the same problem of bound discretion applies in such cases as well.

At first sight, this question seems strange: enemy soldiers are fighting the state, and therefore it is unreasonable to argue that there are any administrative-trusteeship relations between the state and troops fighting against it. However, this intuition is not sufficient, since the mere fact that an individual, say, a known terrorist, threatens state security, does not in and of itself negate the state's administrative obligations towards him or her.

Therefore, in order to answer the question, we need to say something about the nature of the moral justification to target enemy soldiers during war. Chiefly, the question is whether the morality of targeting is status-based or threat-based. On the one hand, we might say that soldiers are *per se* targetable by virtue of their legal status. Indeed, this is the way positive law has traditionally been interpreted.[30] If this is true, there is no need to exercise substantive discretion when deciding to target them. However, legal status notwithstanding, the common moral justification for targeting combatants is still constructed around a notion of threat: meaning that combatants can be attacked because they are presumed to be threatening.[31] If we admit that threat plays any part in the moral justification of targeting – and even if we concede that a strong presumption of threat exists when combatants are involved – this must mean that some kernel of discretion must remain throughout. For instance, the fact that combatants rendered *hors de combat* cannot be targeted requires exercising discretion in determining whether, in a specific case, a person is indeed *hors de combat*.

29 Schmitt and Thurnher (n 26), p. 266.

30 See e.g. G. Blum, The Dispensible Lives of Soldiers, 2 Journal of Legal Analysis 115 (2010) 123-126.

31 M. Walzer, Just and Unjust Wars, 4th ed., New York 2006, p. 145; J. McMahan, Killing in War, Oxford 2011, pp. 32-37.

In the same vein, in recent years there have been significant challenges – albeit still, perhaps, *de lege ferenda* – to the traditional view that combatants are *per se* targetable. At first, some authorities have pointed to a requirement that civilians, even if directly participating in hostilities, must be arrested whenever possible – and killed only if impossible.[32] The kill/capture debate then migrated to the question of whether the rights of *soldiers* also spawn a 'duty to capture'. In this context, some scholars press for the recognition of such an obligation, whether through a narrow interpretation of the concept of military necessity or through an expansive understanding of the notion of *hors de combat*.[33] Whatever the basis for possible obligations to prefer the capture of enemy combatants, their underlying assumption correlates with diagonal, administrative-like obligations between a state and enemy combatants *qua* individuals. To the extent that these perceptions gain traction, they will of course strengthen the need for substantive discretion even in 'pure', symmetrical battlefields, where civilian casualties are unlikely.

It must be emphasised furthermore that the claim that in such 'pure' situations AWS would be legal notwithstanding their inability to exercise discretion, assumes, in fact, that such sterile environments do exist nowadays (or are at least common enough to make a difference).

However, as aforementioned, most modern conflicts tend to be asymmetric and occur in complex environments. Moreover, in asymmetric conflicts, significant discretion is required in order to ascertain whether a person is a combatant, or otherwise directly participating in hostilities to begin with.[34]

In sum, it seems that administrative legal obligations – chiefly the duty to exercise discretion – might apply even vis-à-vis enemy troops. If this is indeed the case, the problem of bound discretion, as posed by AWS in other contexts, applies to enemy troops as well.

32 Case HCJ 769/02, The Public Committee against Torture in Israel v. The Government of Israel (Public Committee v. Israel), 62(1) PD 507, para. 40 [2006] (Isr.); N. Melzer, Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law, Geneva 2009, part IX; but see W. Hays Parks, Part IX of the ICRC 'Direct Participation in Hostilities' Study: No Mandate, No Expertise, and Legally Incorrect, 42 NYU Journal of International Law and Politics 769 (2010) 783-785.

33 R. Goodman, The Power to Kill or Capture Enemy Combatants, 24 European Journal of International Law 819 (2013); but see M.N. Schmitt, Wound, Capture or Kill: A Reply to Ryan Goodman's "The Power to Kill or Capture Enemy Combatants", 24 European Journal of International Law 855 (2013).

34 See e.g. the elaborate debate in Melzer (n 32).

The challenge of 'dumb' time-suspended weapons

One key challenge must be addressed. It is arguable that 'dumb' kinetic weapons such as bullets, ballistic rockets, or artillery rounds also constitute cases of bound discretion. This is because once fired, there is no turning back, thus discretion is terminated upon the weapon's discharge. However, this *absurdum* challenge is unconvincing since the time gap between the exercise of human discretion and the weapon's impact is negligible in such cases. Change of circumstances between release and impact, of the type that requires the reengagement of discretion, is highly unlikely. AWS, on the other hand, would probably be designed to act independently for prolonged periods of time, in which circumstances might change significantly, and thus require renewed human discretion. Moreover, such dumb weapons do not presume to exercise discretion to begin with – rather, they merely follow simple physical rules, and therefore, human discretion upon launch is generally sufficient.

A related argument is that some 'dumb' weapons such as landmines indeed exhibit a time gap between deployment and impact, just as AWS. However, this claim does not vindicate AWS as much as it exposes the problems with landmines and similar weapons: they are perhaps indiscriminate because they involve binding of discretion. Moreover, the 'dumbness' of landmines will be taken into consideration by the reasonable commander and is likely to restrict their use. Conversely, the perceived sophistication of AWS, and their chimera of discretion, will achieve just the opposite: the commander will regard this 'discretion' as sufficient from the perspective of the law. As opposed to the case of landmines and similar 'dumb' weapons, the commander will be more inclined to absolve herself from exercising discretion by relying on the weapon's discretion instead.

Conclusion

Much of the current debate on AWS refers to the need for 'meaningful human control' or 'appropriate levels of human judgment'. Such terms are not self-explaining and require significant interpretation. As this chapter has argued, human control or judgment must be understood in the context of modern warfare, which is closer to executive-administrative action than to classic wars. Since AWS do not exercise substantive discretion – because they cannot engage in metacognition and are based on predetermined algorithms – their use amounts to executive action based on bound discretion. Meaningful human control or appropriate judgment must thus be understood to imply that human beings – as the only agents of 'true' discretion –should make final targeting decisions when human lives are affected.

Meaningful Human Control

Richard Moyes*

Overview

'Meaningful human control over individual attacks' is a phrase that was coined by the NGO Article 36, to express the core element that is challenged by the movement towards greater autonomy in weapons systems.[1] The central area of concern regarding the development of autonomous weapons systems (AWS) is that they might lack the necessary human control in the critical functions of identifying, selecting, and applying force to targets. Without the necessary human control, such systems might not allow the proper application of legal rules, or might produce interpretations of the legal framework that erode civilian protection, or lead to other negative outcomes relating to the morality of human interactions or the maintenance of peace and stability.

In this context, this paper argues that:

- consideration of the form and nature of human control considered necessary is the most useful starting point for discussions on this issue;
- the existing legal framework of international humanitarian law provides a framework that should be understood as requiring human judgment and control over individual 'attacks' as a unit of legal management and tactical action;
- that without recognising a requirement for human control to be in some way substantial or meaningful, the existing legal framework does not ensure that human legal judgment will not be diluted to the point of being meaningless, as a result of the concept of 'an attack' being construed more and more broadly;
- against that background, delineation of the key elements of human control should be the primary focus of work by the international community;
- towards such a process, the following key elements can be proposed:
  - predictable, reliable and transparent technology;
  - accurate information for the user on the outcome sought, the technology, and the context of use;
  - timely human judgment and action, and a potential for timely intervention;
  - accountability to a certain standard;

* Managing Partner at Article 36 (www.article36.org).

1 See for background http://www.article36.org/publications/#kr.

- whilst consideration of these key elements does not provide immediate answers regarding the form of control that should be considered sufficient or necessary, it provides a framework within which certain normative understandings should start to be articulated, which is vital to an effective response to the challenge posed by autonomous weapons systems;
- an approach to working definitions based on understanding 'lethal autonomous weapons systems' as weapons systems operating with elements of autonomy and without the necessary forms of human control would be the most straightforward way to structure discussion in a productive normative direction.

'Meaningful human control' is a policy formulation that has been picked up and used in different ways by different actors – in publications by various individuals and organisations, in state interventions at the UN Convention on Certain Conventional Weapons (CCW), in the open letter from artificial intelligence (AI) practitioners organised by the Future of Life Institute. As used by Article 36, it has always been presented as an approach for structuring a productive debate rather than as providing a conclusion to that debate.

Asserting a need for meaningful human control is based on the idea that concerns regarding growing autonomy are rooted in the human aspect that autonomy removes, and therefore describing that human element as a necessary starting point if we are to evaluate whether current or future technologies challenge that. This is particularly important if a coherent policy conversation is to be had about diverse and often hypothetical future technologies. It is also a starting point for policy that is arguably more open to engagement from diverse stakeholders that might have different expectations of the advantages that may be afforded to them by future developments in autonomous weapons systems.

Considering the key elements necessary for human control to be meaningful does not preclude consideration of other more specific issues – but a structured analysis tends to find that those more specific issues fall within the key elements of human control. For example, the need for 'predictable' technology, the need for human 'judgment' to be applied in the use of force, and the need for accountability all fall under the key elements of human control as laid out in this chapter. Furthermore, without a normative requirement regarding human control, the legal framework itself is open to divergent and progressively broader interpretations that may render human legal application meaningless.

Recognising the need for human control in some form

At its most basic level, the requirement for meaningful human control develops from two premises:

1. that a machine applying force and operating without any human control whatsoever is broadly considered unacceptable;
2. that a human simply pressing a 'fire' button in response to indications from a computer, without cognitive clarity or awareness, is not sufficient to be considered 'human control' in a substantive sense.

On this basis, some human control is required and it must be in some way substantial – we use the term 'meaningful' to express that threshold. From both of these premises, questions relating to what is required for human control to be 'meaningful' are open. Given that openness, meaningful human control represents a space for discussion and negotiation. The word 'meaningful' functions primarily as an indicator that the form or nature of human control necessary requires further definition in policy discourse.

Critical responses to this policy formulation tend to fixate on the term 'meaningful' because it is undefined or might be argued to be vague – responses that may also be motivated by state representative anxieties at policy formulations not initiated by states. Such responses, however, miss the point. There are other words that could be used instead of 'meaningful', for example: appropriate, effective, sufficient, necessary. Any one of these terms leaves open the same key question: how will the international community delineate the key elements of human control needed to meet these criteria? Any one of these would also be vague until the necessary form of human control is further defined, giving the chosen adjective some further calibration.

The term 'meaningful' can be argued to be preferable because it is broad, it is general rather than context-specific (e.g. appropriate), derives from an overarching principle rather being outcome-driven (e.g. effective, sufficient), and it implies human meaning rather than something administrative, technical, or bureaucratic.

That said, fixating on which adjective is most appropriate should not stand as a barrier to the next step required of the international community, which is to begin to delineate the elements of human control that should be considered necessary in the use of force.

Situating human control in the legal framework

The NGO Article 36 has called on states, in the context of discussions on autonomous weapons systems in armed conflict, to recognise the need for 'meaningful human control over individual attacks'. In its use of the term 'attacks', this formulation situates the issue of human control within the legal framework of international humanitarian law (IHL).

It is important to recognise that IHL is not the only legal framework relevant to AWS, nor are legal frameworks the only basis for assessing whether further development of such technologies is appropriate or advisable. However, the relationship between human control, AWS, and IHL is given particular focus in this chapter.

Human beings as addressees of the law

When discussing AWS, however complex, the NGO Article 36 refers to these systems as 'machines'. Discussion on this issue is prone to a slippage towards treating these machines as 'agents' and in particular as 'legal agents'. It is common for diplomats and experts to refer to concerns about whether AWS will "be able to apply legal rules", or "to follow the law". Machines do not apply legal rules. They may undertake functions that are in some ways analogous to the legal rules (for example being programmed to apply force to certain heat patterns common to armoured fighting vehicles) but in doing so they are not "applying the law" – they are simply implementing a process that a human commander anticipates in their assessment of the legality of a planned attack. Marco Sassòli, in his presentation to the 2014 ICRC expert meeting on autonomous weapons, stated that "only human beings are addressees of international humanitarian law".[2]

2 M. Sassòli, Can Autonomous Weapon systems Respect the Principles of Distinction, Proportionality and Precaution?, ICRC, Report on the Expert Meeting – Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects, 26-28 March 2014, https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf.

Human judgment in relation to 'attacks' – part of the structure of IHL

Given that human beings are the addressees of the law, whether collectively or individually, there are certain boundaries of machine operation that the law implies in relation to humans. The term 'attacks' in IHL provides a unit of military action, and it is over individual 'attacks' that certain legal judgments must be applied. So attacks are part of the structure of the law, in that they represent units of military action and of human legal application.

For example, Article 57 of Additional Protocol I (API) provides rules on precautions to be taken in attack. Where it refers to "those who plan or decide upon an attack", it is referring to humans. It is therefore humans that shall apply these legal rules – including verifying the objective, choosing the means and methods of attack, and refraining from or cancelling an attack in certain circumstances.

We know that an attack must be directed at a specific military objective, as otherwise it is indiscriminate (Article 51(4)(a) API). We also know that a military objective must be of a sort (nature, location, etc.) to offer military advantage at the time (Article 52(2) API), and that in the application of the legal rules the concrete and direct military advantage must be assessed by the humans that plan and decide upon an attack (Article 51(5)(b) and Article 57(2)(a)(i) and (ii) API). Therefore, humans must make a legal determination about an attack on a specific military objective based on the circumstances at the time. There should also be a capacity to cancel or suspend an attack (Article 57(2)(b) API).

These rules imply that a machine cannot identify and attack a military objective without human legal judgment and control being applied in relation to an attack on that specific military objective at that time (control being necessary in some form to act on the legal judgment that is required). Arguing that this capacity can be programmed into the machine is an abrogation of human legal agency – breaching the 'case-by-case' approach that forms the structure of these legal rules.

This line of argument is not dependent upon claims regarding the technical capacity of complex future AWS to do this or that, but is based on the law as a framework that applies to humans and that is structured to require human legal judgments at certain points.

However, this is not to argue that the law straightforwardly implies a very narrow constraint on what an AWS might do under its existing terms. Nor is it suggesting that existing law alone represents a sufficient basis for managing AWS. It is simply to point out that the existing legal structure (human judgment being required over 'attacks') implies certain

boundaries to independent machine operation and that this is separate from arguments about how a machine might perform in relation to the implementation of individual legal rules (for example, the rule of proportionality).

Conceptualising 'an attack'

Whilst seeing in the structure of the law an assumption of human legal judgment in relation to individual attacks, it is also recognised that 'an attack' is not necessarily a single application of kinetic force to a single target object. In practice, an attack may involve multiple kinetic events against multiple specific target objects. However, there have to be some spatial, temporal, or conceptual boundaries to an attack if the law is to function. This is linked to the different layers at which military action is often conceptualised – from the local tactical level, through the operational to the broad strategic level. If 'attacks' were not conceptualised and subject to legal judgment at the tactical level, but only, for instance, the broad strategic level, then a large operation may be determined to be permissible (on the basis of broad anticipated outcomes) whilst containing multiple individual actions that would in themselves be legal violations. Clearly, for the law to function meaningfully, there needs to be legal judgments and accountability over actions at the most local level.

Recognition that human legal engagement must occur over each attack means that a machine cannot proceed from one attack to another, to another, without human legal judgment being applied in each case, and without capacity for the results of that legal judgment to be acted upon in a timely manner – i.e. through some form of control system. Given that an attack is undertaken, in the law, towards a specific military objective that has been subject to human assessment in the circumstances at the time, it follows that a machine cannot set its own military objective without human authorisation based on a human legal judgment.

Preventing an expansion of the concept of 'an attack'

Our starting point in this chapter was the concern that greater autonomy in weapons systems may result in human control not being meaningful. Based on the analysis above regarding the relationship of autonomy to the legal framework, we can see that this concern is linked to a risk that autonomy in certain critical functions of weapons systems might produce an expansion of the concept of 'an attack' away from the granularity of the tactical level, towards the operational and strategic. That is to say, AWS being used

in 'attacks' which in their spatial, temporal or conceptual boundaries go significantly beyond the units of military action over which specific legal judgment would currently be expected to be applied.

Greater specificity of legal assessment – by this we mean a legal assessment that is evaluating specific events expected to occur over a shorter period of time, and within a narrower area – allows for specific risks to the civilian population to be more accurately assessed, and therefore for civilians to be better protected. Furthermore, allowing greater autonomy to facilitate progressive broader interpretations of what constitutes an attack would have a corrosive function upon the legal framework as a whole. This raises a key objection to assertions that national weapon review processes would be a sufficient response to the concerns posed by autonomous weapons. If the very tests that are applied to determine permissibility of a weapons system are being undermined by the development of that weapon system itself, how can the review process remain meaningful?

By asserting the need for meaningful human control over attacks in the context of autonomous weapons systems, states would be asserting a principle intended to protect the structure of the law, as a framework for application of wider moral principles. Moving the debate on to delineate the elements needed for human control to be meaningful would start to develop a normative understanding that should pull towards greater granularity and specificity of legal assessment, rather than greater generalisation.

Key elements of human control

So, as framed by the previous section, a meaningful form of human control is necessary both to allow for legal application and to protect the structure of the law from progressive erosion. In that context, the section below sketches out 'key elements' through which human control can be understood to be applied in the use of weapons systems. These elements are not simply about technological characteristics but recognise that human control is necessarily part of a wider system that allows a specific technology to be controlled in a specific context of use.

Predictable, reliable, and transparent technology

Starting with the technology itself, human control is facilitated where the technology is:

- predictable – it can be expected to respond in certain ways;
- reliable – it is not prone to failure, and is designed to fail without causing outcomes that should be avoided;
- transparent – practical users can understand how it works.

In whichever way the technology is to be used, there are certain characteristics that may be designed and manufactured into the technology that have a bearing upon the subsequent capacity for human control. A technology that is by design unpredictable, unreliable and un-transparent is necessarily more difficult for a human to control in a given situation of use.

Accurate information for the user on the outcome sought, the technology, and the context of use

Human control in the use of a technology is then based upon those planning and deciding upon an attack having certain information. Control in the use of a weapons system can be understood as a mechanism for achieving commander's 'intent'. So information on the objective that is sought is an important starting point – including information on the unintended consequences that a commander wishes to avoid. This information is necessary for a human commander to assess the validity of a specific military objective at the time of an attack, and to evaluate a proposed attack in the context of the legal rules.

Such assessments also require an understanding of the technology. For example, we need to know what types of object a weapons system will identify as a target object – target 'profiles' – whether these are the commander's intended targets or not. We need to know how kinetic force will be applied – it makes a difference if the force will be a heavy explosive weapon with large blast and fragmentation radius, or if it will apply force quite narrowly, such as with an explosively formed projectile with no fragmentation effects.

'Predictability' is an important concept in that it provides a link between commander's intent and the likelihood of outcomes that match that intent. Predictability is partly a characteristic of the technology, but more fundamentally it is a characteristic of the interaction between that technology and the specific environment within which it will operate. As a

result, information on context of use is very significant. We should have some understanding of the environment in which the technology will operate, including the presence of civilians and civilian objects, for example.

Of course, we may not achieve complete predictability – already in the use of weapons we accept degrees of uncertainty about the actual effects that will occur, and we know that there may be limitations on the information available about the context. However, our ability to understand the context is directly linked to both the size of the area within which the technology will operate, and the duration over which it will operate. For any given environment, it follows logically that greater area and longer duration of independent operation by a technology result in reduced predictability, and therefore reduced human control.

It is recognised that different environmental domains present different general characteristics – with land, air, and sea presenting different levels of complexity. This may mean that a large area of operation in the sea may still facilitate better contextual understanding than a smaller area on land. However, for environments of equal complexity, greater area and greater time of operation necessarily mean reduced control. In relation to the duration of an attack, this might be because certain people or objects enter or leave an area over time in a way that could not be anticipated, or it could be because the commander's intent has changed from the point at which the attack was initiated.

From an understanding of the technology, and an understanding of the context within which it will operate, a commander should able to assess likely outcomes, including the risk of civilian harm, which is the basis for the legal assessment. It is important to note that information on these different elements may be the product of wider human and technological systems, but at some point understanding of these three elements must coalesce to a point where an informed judgement can be made.

Timely human judgment and action, and a potential for timely intervention

Based on the information on the outcome sought, the technology and the context, we need humans to apply their judgment – as implied by the legal analysis earlier in this paper – and to choose to activate the technology. This point of human engagement ties together the systems of information upon which judgments are made, but also provides a primary point of reference for the framework of accountability within which these actions are taking place. Of course, responsibility for negative outcomes may turn out to result from problems

elsewhere in the system (e.g. malfunctioning technology or inaccurate information on the context of use), but human judgment and action at this point is likely to be the starting point from which any negative outcomes are investigated.

The timeliness of this process is also significant because the accuracy and relevance of the information upon which it is based, for example about context, also degrades over time. For a system that may operate over a longer period of time, some capacity for timely intervention (e.g. to stop the independent operation of a system) may be necessary if it is not to operate outside of the necessary human control.

A framework of accountability

Finally, this broad system requires structures of accountability. Such structures should encompass not just the commander responsible for a specific attack, but also the wider system that produces and maintains the technology, and that produces information on the outcomes being sought and the context of use.

Conclusion on the key elements of human control

All of these areas cumulatively contribute to the extent of human control that is being applied in a specific context of use. In all of these areas there are tests of 'sufficiency' that would need to be met in order for the overall extent of human control to be assessed as sufficient in itself. Where some have asserted that the existing legal framework provides the answers needed for evaluating autonomous weapons systems, these tests suggest that this is not straightforwardly the case.

For example, it is not clear what level of information about the context within which a weapon will be used is considered 'sufficient' to provide a basis for an informed legal judgment. If a weapons system were to apply force to the individual vehicles of a group of fighting vehicles, this might be considered reasonable if the group were known to be in a reasonably bounded geographical area over which a commander had knowledge. However, if the area within which that group of vehicles was situated was spread over a wider area, about which the commander necessarily had a lesser and lesser understanding, at what point does that understanding become so diluted as to make a legal assessment unreasonable? In legal terms, this is a question about what can reasonably be considered a 'specific military objective' and about what can reasonably be considered 'an attack'. The law alone

does not provide an answer to these questions that resolve the uncertainty here, yet such questions are fundamental to avoiding the erosion of the legal framework that can be envisaged should states choose to develop autonomous weapons systems.

Whilst consideration of the key elements of human control does not immediately provide the answers to such questions either, it would at least allow states to recognise that these questions are fundamental, and it provides a framework within which certain normative understandings should start to be articulated, which is vital to an effective response to the challenge posed by autonomous weapons systems.

Working definitions – facilitating discussion within the CCW

The most direct way in which to establish such a discussion within the CCW is to adopt an approach to working definitions that is based on a recognition that certain forms of human control are required over the use of force, and that systems operating outside of that definition should not be considered acceptable. That would most straightforwardly be facilitated by adopting a working definition of 'lethal autonomous weapons systems' that is based on these being 'weapons systems operating with elements of autonomy and without the necessary forms of human control'. In such an approach, the concept of weapons systems operating with elements of autonomy then refers to a broad category of systems within which a certain subset (either by design or by their manner of use) is considered unacceptable. Such an approach then sets up delineation of the key elements of human control as a primary focus of work in order to understand where the boundaries of permissibility should lie.

Approaches to Legal Definitions in Disarmament Treaties

Gro Nystuen*

Introduction

In international law and treaty making, the 'question of definitions' is often crucial. Aiming to facilitate discussion on how a legal definition of autonomous weapon systems might be constructed, this chapter looks at how specific weapons have been defined in international disarmament treaties.

Before proceeding, though, it is worth noting that not all weapons regulated by an international instrument are explicitly or exhaustively defined in the instrument itself. For example, neither the Nuclear Non-Proliferation Treaty[1] (NPT) nor the Comprehensive Nuclear Test-Ban Treaty[2] (CTBT) define nuclear weapons. Similarly, the Biological Weapons Convention (BWC) does not define biological weapons. Accordingly, an explicit definition is not necessarily a precondition for agreeing on measures to regulate a category of weapons.

The treaties that do contain explicit definitions highlight different aspects of the weapons in question. However, all of them contain a reference to the larger category or family of objects to which the weapon belongs. An anti-personnel mine is a type of 'mine'; a cluster munition is a type of 'conventional munition'; a 'booby-trap' is a type of 'device or material' and so on.

* Senior Partner, International Law and Policy Institute, Oslo.

1 Treaty on the Non-Proliferation of Nuclear Weapons (NPT), opened for signature on 1 July 1968.

2 The Comprehensive Nuclear Test-Ban Treaty (CTBT), opened for signature on 24 September 1996.

Operation

In distinguishing the weapon in question from its broader category, disarmament treaties have applied different methods and parameters. A common 'distinguishing attribute' is the intended or actual functioning or operation of the weapon. For example, the Convention on Cluster Munitions' (CCM)[3] definition of a cluster munition contains a description of how cluster munitions are delivered. The Anti-Personnel Mine Ban Convention[4] (APMBC) and Protocol II of the Convention on Certain Conventional Weapons (CCW)[5] define mines by describing how these munitions are detonated. A key feature of the definition of anti-personnel mines is that they are *victim-activated*. If the mine functions through manual or remote detonation, i.e. by being set off by someone other than the victim, it falls outside the definition of an anti-personnel mine. The defining difference between anti-personnel mines, on the one hand, and anti-vehicle mines, on the other, is that while the former are set off by persons, the latter are triggered by vehicles.

Effects

Another attribute used to define a category of weapons is their effects. CCW Protocol I on blinding lasers, for example, says nothing about how the weapons function, but highlights instead the intended or actual effect of the weapons, which is permanent blindness. Protocol III in the CCW describes incendiary weapons as weapons that cause burn injuries to persons. Other treaties refer to both intended functioning and effects: an anti-personnel mine is a mine designed to "incapacitate, injure or kill", a chemical weapon is a toxic chemical leading to "death, temporary incapacitation, or permanent harm".

3 The Convention on Cluster Munitions (CCM), opened for signature on 3 December 2008.

4 The Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction (APMBC), opened for signature on 3 December 1997.

5 The Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (CCW), opened for signature on 10 October 1980.

Targets and victims

A third distinguishing attribute sometimes applied in definitions is the weapon's intended or unintended targets or victims. While the Mine Ban Convention refers to "one or more persons", the Chemical Weapons Convention (CWC) refers to "humans" or "animals". ENMOD,[6] for its part, refers to the natural environment.

Intended use

A fourth distinguishing attribute that might be especially relevant in the context of autonomous weapons concerns the intended use of the weapon. This concerns in particular weapons where there is dual-use potential. Weapons with limited dual-use potential, such as cluster munitions, do not have — or, indeed, require — a reference to the intended use of the weapon. Weapon categories with more pronounced dual-use potential, however, such as chemical and biological agents, do contain this qualifier.

For example, the CWC defines chemical weapons as a toxic chemical that, "through chemical action", can temporarily or permanently incapacitate or kill human beings or animals. This definition is very broad, and would seem to include for example chemicals such as aspirin or alcohol. Consequently, the exceptions also have to be extensive. The CWC determines that chemical agents intended to be used for industrial, agricultural, research, medical, pharmaceutical or other peaceful purposes; as well as protective purposes, or military purposes not connected with the use of chemical weapons, are all explicitly exempt from the definition. The same applies to chemical agents used for law enforcement purposes including domestic riot control. The prohibition on chemical weapons thus encompasses use of chemical agents as means of warfare. Biological weapons, moreover, while not defined in the BWC,[7] are implicitly understood as a living organism (which is a very broad term) used "for hostile purposes".

6 The Convention on the Prohibition of Military or Any Other Hostile Use of Environmental Modification Techniques (ENMOD), opened for signature on 18 May 1977.

7 The Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction (BWC), opened for signature on 10 April 1972.

Linkages to international humanitarian law

Arms treaties sometimes refer to rules under international humanitarian law (IHL) when defining the scope of the treaty. The most obvious case is the CCW Framework Convention, which makes it clear through its title that the protocols are directly relevant for, and aim at strengthening the implementation, of two fundamental rules under IHL: the prohibition against means and methods of warfare that lead to superfluous injury or unnecessary suffering for combatants, and the prohibition on means or methods that are indiscriminate, thus violating the rule of distinction. In the CCW protocols, these qualifiers are not explicitly included in the various definitions, but it is still clear that this reference to two core rules under IHL is the very basis for the protocols.

In the CCM, the definition of cluster munitions contains a direct reference to IHL. Its article 2(c), where the exceptions to the wider category are listed, starts with: "A munition that, in order to avoid indiscriminate area effects and the risks posed by unexploded sub-munitions, …". These qualifiers pertain directly to the rule on distinction.

Scope of application

One key feature in disarmament treaties, which is not strictly a part of the definitions of the weapon, but which nevertheless is very interesting with regard to autonomous weapons, is the scope of application, or in other words: in what situations does the treaty apply? The CCW has a relatively limited scope compared to many other arms treaties. Originally, it applied only in international armed conflicts. It was later amended to include also non-international armed conflicts,[8] but not all State Parties to the CCW have accepted this amendment, and thus the protocols only apply in international armed conflict for those states.

Other disarmament treaties, on the other hand – the BWC, the CWC, the APMBC, and the CCM, for example – apply in *all* circumstances, regardless of the classification of the situation, and so they apply also in situations that do not reach the threshold of armed conflict.[9] The issues regarding whether autonomous weapons could be used outside of an armed conflict, for law enforcement purposes, will be a key point of contention. If the debate on autonomous weapons develops into a process towards a protocol under the CCW, the issue of the scope of application will have to be addressed.

8 Art. 1 (2) CCW, amended on 21 December 2001.
9 The CWC has specific exceptions regarding the use of toxic chemicals in law enforcement.

Is a precise definition a prerequisite for negotiations?

A final point worth mentioning when discussing definitions in disarmament treaties is about process rather than the substantive elements of a definition. Most international legal processes aiming to ban or regulate specific weapons have started and proceeded without agreement on the definition of the object(s) to be regulated or banned. Key definitions are typically among the very last issues to be settled in negotiations. This was the case, for example, in the negotiations of the Chemical Weapons Convention (adopted in 1993), the Mine Ban Convention (adopted in 1997), and the Convention on Cluster Munitions (adopted in 2008). To have an agreed definition is rarely a prerequisite for proceeding with negotiations on arms regulations.

All weapons, including autonomous weapons, have to comply with IHL's rules on means and methods of warfare. One of the key discussions with regard to autonomous weapons will be whether autonomous weapons fall into the broader category of "means" (weapon systems) or "methods" (how weapons are used), or both.

Autonomy in Weapons Systems: Past, Present, and Future

Heather M. Roff*

Introduction

When one considers the current state of affairs in regard to autonomous capacities in weapons systems, it is imperative to ground one's estimations in empirical evidence. My contribution on 'mapping autonomy' attempted to do this by utilising data to answer four main questions: (1) What is the state of military weapons technology today; (2) Where do we see autonomy in 'critical functions'; (3) What is the trajectory of autonomy in weapons systems; (4) Where will we likely see autonomous weapons develop?

The state of military weapons technology today

To assess the present state of military weapons systems, I generated a data set of presently fielded weapons systems. This was restricted to the top five weapons exporting countries (the United States, Russia, China, Germany, and France). Moreover, for the sake of ease of comparison, I only surveyed their presently deployed missile and bomb arsenals. Choosing to focus solely on missiles and bombs allows us to see trends of developments over time, as we have ample data on past and legacy systems, as well as systems that are acquired in 'blocks'. In other words, there may be older models of a missile, and through various 'blocks' (block 1, 2, 3, etc.) there are upgraded capacities added to the system. Moreover, the choice of these five countries was to account for the majority of the world's arms development and trade. These five countries make up 74 percent of the world's arms trade, and as such are leaders in weapons development and export. Finally, the data set consists of over 230 weapons systems.

The data suggests that most advancements relate to homing, navigation, target acquisition, target identification, target prioritisation, auto-communication, and persistence (or the ability to loiter). Systems are able to direct themselves to particular locations in space or to particular targets, and once there, more advanced systems can identify targets automatically or may be able to communicate with other deployed munitions. Present day systems lack

* Senior Research Fellow, Department of Politics & International Relations, University of Oxford; Research Scientist, Arizona State University.

the ability to give themselves goals or missions, and only some systems are able to update or change plans once deployed. The ability to change plans is most often related to navigation functions and not to the prosecution of attack.

Autonomy in 'critical functions'

Autonomy in 'critical functions', or those functions related to the selection and engagement of a target, is present in some current systems. However, there is open debate as to whether 'autonomy' here means the mere ability to respond or react without intervention by a human or direction by a human operator, or something more robust, such as greater cognitive capacities in making a 'decision'. In mapping the current state of weapons deployment, I took a neutral stance with regard to definition. To ensure this, I merely identified all the relevant capabilities on a system (e.g. homing, navigation, mobility, automatic target recognition, etc.), and then I coded each variable as binary (the system either has it (1) or it does not (0)). For example, there are systems that possess automatic target recognition software, enabling them to find a target on their own, match that target to a target identification library or database, and then fire on the target. This is coded as a (1). What is more, close-in defensive weapons systems are also capable of sensing a target, prioritising that target, and firing on it without the intervention of a human operator, these are also coded as a (1). This binary or 'dummy variable' system allows me to move away from assessments of whether the system is 'autonomous' or 'automatic', or in attempting to fit a particular system on any sort of scalar or 'level' approach.

That said, in current weapons systems, the 'selection' of targets may be better thought of as 'detection'. Present-day systems have various sensor capabilities that allow them to perceive the environment around them, and then to recognise potential targets (such as enemy radars or tanks). Once deployed, these systems are constrained in the types of targets they can fire upon, as only those targets that match the target identification library would be seen as 'matches'. In cases where a specific location in space is the target area, that location has been chosen by a human, and in cases where lasers are designating a target object, a human is also choosing that target. In the limited cases, such as anti-ship missiles, these systems are also utilising various sensor capacities to navigate, locate and identify targets (ships). Once there, they are able to select amongst various identified targets, but it appears that they do so by prioritising detected objects, ostensibly given some sort of predefined criteria.

Trajectory of autonomy in weapons systems

The trajectories of autonomy in weapons system can be considered along several continuums. From the 1960s onwards, there were significant developments in homing, navigation, and mobility. Instead of dropping unguided bombs, developments towards self-propelled guided missiles were of primary importance for reasons ranging from force protection, power projection, and civilian safety.

The 1970s to 1980s began to see more development in capabilities related to target identification, image discrimination, and target ranking or prioritisation. These advancements are more than likely due to the technological advances made in sensor technologies in the 1970s, as well as advancements in image processing capabilities, such as through software development, microelectronics, and microprocessor speeds in the 1980s. What is more, the continued pursuit of long-range munitions required that they be able to direct themselves to particular targets and, once there, to identify those targets. Cruise missiles, for example, are 'fire and forget' weapons, in that once they are launched they will navigate themselves to a particular location in space to detonate. Thus, strategic choices related to military strategy and policy, including standoff capabilities, affected acquisition and adoption of more self-mobile and self-directed weapons.

Today, with advances in machine learning, especially related to image recognition and classification, there are movements to utilise these technologies in target recognition. Particularly, there is desire to use advances in artificial intelligence to enable automatic target recognition to adapt and learn new targets when an adversary force changes tactics. With the vast amount of processing power we presently possess, there are also additional attempts to reconfigure hardware to make them more efficient in their power and processing abilities. In effect, this allows previously intractable computing problems to be made possible in short time with low power consumption. Moreover, with growing capabilities to deny manoeuvrability or use of standoff weapons, militaries are also seeking to find new ways of utilising miniaturisation in electronics and robotics. Progress in swarming techniques is also enabling autonomous capacities in groups of vehicles or vessels so that these systems will be able to prosecute attacks with or without direct communication links.

Areas of autonomous weapons development

There are potentially three areas to consider for autonomous weapons systems development: single platforms, combinations of legacy systems, and modular systems.

Single platforms

Single platform weapons systems or munitions, such as missiles, bombs, torpedoes, or mines, are one potential area of autonomous weapons development. Such systems are better thought of as either a single platform (or swarm) with munitions on-board, or as a single munition (such as a loitering munition). We might want to consider the development of these types of unitary autonomous weapons as *intentional* autonomous weapons. These are likely to be used in conjunction with other systems or platforms, but they can be thought of as a 'closed' or unitary. Maritime and air domains are the most likely domains in which these systems would be used, as there are fewer difficulties with obstacle avoidance.

Combinations of legacy systems

Nevertheless, there is a likelihood that autonomous weapons systems will not first appear in the form of single platform or unitary munitions. Rather, what is more likely is the combination of various legacy systems that enable a *functionalist* approach to autonomous weapons systems. In other words, depending upon the type of task or mission requirement, militaries may combine existing unmanned platforms with one another in collaborative exercises to yield a functionally autonomous weapons system. Air, land and sea may be combined into one large system, with various semi-autonomous and/or loitering munitions attached to these platforms. The result would be that human control over critical functions may be stressed or functionally eliminated, whereby the actual choice over targets is not under the control of a human operator or commander.

Instead, a human commander chooses the battlespace, and any potential targets within that space are 'selected' by the weapons systems (e.g. if the battlespace is suppressing enemy air defences, etc.). The human commander cannot know which targets will be destroyed, except that they will be in a particular geographic area. Depending upon the autonomous capacities of the platforms (such as mobility, navigation, auto-communication sharing, etc.); number of platforms in the collaborative operation; geographical space with which the systems can function; time that such systems can operate or extend operations through

deployment of further loitering sub-munitions, one could judge that though no one single platform is an 'autonomous weapon', the combination of multiple semi-autonomous systems yields an autonomous weapons system in a larger and functionalist sense.

Modular weapons systems

Unlike the scenario above, where existing platforms and munitions are combined to yield a functionally autonomous weapons system, the modular approach to autonomous weapons is where various components for platforms, munitions, sensors and the like are produced as stand-alone modular components that can be assembled in various configurations. This approach would entail a blending of the intentionalist and functionalist approaches to autonomous weapons. Here there is no single, unitary, autonomous weapon designed for one role, but neither is there a combination of existing unitary semi-autonomous weapons in a collaborative role that yields a functionally autonomous weapons system. Rather, it is a combination of the two. Each modular component is designed to complete a task and to be compatible with other modular components, with the foreseeability that in certain combinations they may yield autonomous weapons. Such an approach could be domain-specific, such as modular approaches to subsurface systems, or multi-domain, where components may fit on a variety of platforms or munitions spanning air, ground, and sea.

Cyber weapons

While the discussion of autonomous cyber weapons is not within the purview of the current debates in the Convention on Conventional Weapons, there is concern that one of the first areas of autonomous weapons development will be the cyber domain. This domain does not require the pairing of software with robotic systems and thereby decreases the engineering difficulties of integrating large platforms, munitions, and software architectures. Given the vast array of attacking and defending in cyberspace, I cannot traverse the possible futures here. Rather, we would want to say that this domain may require autonomous capabilities because of the vast amounts of data flow that a human could never possibly monitor or respond to in real time.

Decision aides

Finally, autonomous weapons may actually not exist in the sense of an identifiable robotic platform, munition, or software package. Future development on decision aides for command and control will require great care in that the decision aides—that is software programs that may feed information, package information, or present courses of action to commanders—do not end up being 'decision makers'. Given the vast amount of information and data that flows from a battlespace to a commander, one can understand the reasons for developing cognitive augmentations or enhancements that enable a human commander to make the right decisions. However, the character and source of information (reliable, viable, corroborated), as well as the presentation or menu of choices or options (risk analysis, cost-benefits, value distinctions), will affect how a human makes a decision. If systems are not designed with this in mind, then there is risk that a human commander is ostensibly held hostage to the decision aid; that is, the decision aid is actually a decision maker, and the human commander is only one more step in the 'weapons system'. The human acts as a mere placeholder for accountability, but by all accounts has deferred judgment to the software program.

Conclusion

The present state of military weapons systems is such that they possess many capacities related to mobility and direction, however they lack higher order cognitive capabilities. Such presently deployed systems cannot adapt or learn. Some systems, however, do possess planning and prioritising capabilities. In the near term, we will see increases in the creative adaptation and configuration of existing systems to extend the reach and ability of semi-autonomous weapons systems. Future systems will likely utilise more robust planning and learning capabilities, and the likelihood that weapons systems will actually expand, in a nested fashion, to be systems of systems of systems. Adaptable sensor technologies will feed data and information to other platforms for processing, and these platforms may perform decision aide-related activities. Collaborative autonomy between various domains and systems is likely, given the need to manoeuvre in denied environments.

What Is 'Judgment' in the Context of the Design and Use of Autonomous Weapon Systems?

Dan Saxon*

Judgment is the combination of knowledge, experience, talent, reflection, intent, and instinct that allows us to assess, and react to, complex situations. Human judgment includes the ability to decide *when* to let a machine operate autonomously – in other words, to forfeit one's judgment, at least temporarily – and when to maintain control over the machine.

In the context of autonomous weapon systems, computers and the artificial intelligence that drives them may be more efficient than humans in deductive tasks such as filtering information and making calculations. Human judgment, however, is superior in situations requiring inductive reasoning. In other words, human judgment is better and faster for decisions that require analysis of context, conditions and circumstances,[1] because humans have greater ability to apply inductive reasoning for creative thinking.[2] In addition, via experience, humans also develop instincts (our 'sixth sense') that often assist them to navigate difficult situations where strict rules may not suffice.[3] Perhaps ironically, the development of 'good' human judgment often requires divergence from absolute values in order to find solutions to value-based problems.[4]

With respect to the law of armed conflict, the targeting rules found in treaty and customary law (such as the duties of distinction, proportionality, and precautions during attack) apply to the use of autonomous weapon systems (like any other weapon system). Professional armies must "expect military commanders employing a system with autonomous

* Leiden University College.

1 Author interview with M. Johnson, 10 June 2014, Leiden, Netherlands.

2 M. Cummings, Man Versus Machine or Man + Machine?, unpublished draft (copy in author's possession), p. 12. Research efforts are underway to mimic human reasoning and judgment processes in machines. One example is KEEL Technology. "KEEL stands for Knowledge Enhanced Electronic Logic", email from Tom Keeley, 2 and 13 June 2014. See Keel Technology for Complex Problems, available at www.compsim.com/.

3 For example, human soldiers can enter an environment and "get a feeling about it" that often is correct. Machines cannot do that. Interview with former U.S. Army Intelligence Specialist Allen Borrelli, The Hague, 15 July 2015.

4 H. Kelsen, What Is Justice?, in: H. Kelsen, What Is Justice? Justice, Law and Politics in the Mirror of Science: Collected Essays, Berkeley 1957, p. 10.

functions to engage in the decision-making process that is required by international humanitarian law".[5] Logically, it is impossible for commanders to *direct* weapons at specific military objectives, as required by Article 51(4)(b) of Additional Protocol I to the 1949 Geneva Conventions (API),[6] without a proper understanding of the weapon. Thus, deployment of autonomous weapons systems without a proper comprehension of how the system works will constitute an indiscriminate attack and be subject to criminal sanction, at least in jurisdictions that recognise the *dolus eventualis* standard for *mens rea*.[7] Moreover, prior to deploying an autonomous weapon,[8] the superior must ensure one of two criteria: (1) once programmed, the artificial intelligence software controlling the autonomous weapon system has the robust capacity to comply with Article 57 of API, or (2) deployment of the autonomous weapon system is itself an expression of a "feasible precaution in the choice of means and methods of attack" within the meaning and spirit of the law.[9] In some ways, therefore, the use of autonomous weapon technologies will require additional exercises of human judgment, rather than less.

The ability to make the difficult value judgments often present in complex proportionality analysis (as well as other precautions in attack) probably presents the greatest cognitive challenge to the lawful operation of autonomous weapon systems.[10] The data-processing strengths of modern computers miss the qualitative ability to assess the competing human priorities of military advantage and the protection of civilians. This capacity for judgment, the

5 Colonel R. Jackson, Panel on 'Autonomous Weaponry and Armed Conflict', Annual Meeting of the American Society of International Law, Washington DC, April 2014.

6 https://ihl-databases.icrc.org/ihl/INTRO/470.

7 M. Schmitt, Remarks during the panel on 'The International Legal Context' at 'Autonomous Military Technologies: Policy and Governance for Next Generation Defence Systems', Chatham House, London, 24 February 2014. Permission to cite provided in email to the author, 15 March 2014.

8 By definition, once the commander deploys an autonomous weapon platform, she may lose her ability to take additional feasible precautions as well as make proportionality judgments. During the Bill Clinton administration, after U.S. armed forces under his command launched automated cruise missiles against the headquarters of Saddam Hussein's intelligence service in Baghdad, President Clinton was aghast to learn that the missiles neither had cameras mounted on them, nor could they be 'turned back' prior to striking their targets. See R. Clarke, Against All Enemies: Inside America's War on Terror, New York, 2004, pp. 82-83.

9 R. Jackson (n 5); Art. 8(2)(b)(iv) of the Rome Statute of the International Criminal Court prohibits attacks where the anticipated civilian injury and damage is 'clearly excessive' to the expected military advantage. No similar provision exists in treaty or customary law that criminalises failures to take feasible precautions under Arts. 57(2)(a)(i) or (ii). I am grateful to Professor Robin Geiß for clarifying this point.

10 M. Schmitt and J. Thurnher, "Out of the Loop": Autonomous Weapon Systems and the Law of Armed Conflict, 4 Harvard Natl. Sec. J. 231 (2013) 266-267; M. Sassóli, Automomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified, 90 International Law Studies 308 (2014) 331-333.

ability to fuse accumulated knowledge, experience, instinct,[11] and 'common sense', resides, for the time being, in the human mind.[12] Given the present state of artificial intelligence, without human-machine teamwork in situations where proportionality evaluations and other value-based decisions are necessary, the deployment of a lethal autonomous weapon system would be illegal pursuant to the targeting rules of international humanitarian law.

Nevertheless, as the technology improves, it is possible to envisage scenarios where an autonomous weapon system can fulfil targeting obligations more successfully than humans.[13] Tests of new 'machine-learning' systems[14] demonstrate that 'machine-learning' artificial intelligence often exhibits better 'judgment' than humans in response to certain situations.[15] Unburdened by stress and fatigue and capable of processing more data, more quickly, than human soldiers, machines – in some situations – will exhibit more 'tactical patience'[16] and, potentially, more accuracy when distinguishing between civilians and combatants. Moreover, an autonomous weapon system, unworried about its own survival, can delay the use of force, thereby reducing doubt about the military or civilian nature of a target.

Similarly, autonomous weapon systems could provide opportunities for greater pre-cautionary measures – including more accurate proportionality analysis – than human soldiers planning and executing an attack. It can also use less force, including non-lethal

11 Reliance on one's natural instincts, of course, can be fallible. See Aristotle, On Rhetoric: A Theory of Civic Discourse (trans. G.A. Kennedy), 2nd ed., Oxford 2007, p. 94.

12 M. Sassòli (n 10), p. 334; ICRC Commentary to Art. 57 API, at 2208, https://www.icrc.org/applic/ihl/ihl.nsf/INTRO/470.

13 M. Sassòli (n 10), pp. 310-311.

14 Although algorithm-based artificial intelligence is the most common form in use today, 'Statistical Machine Learning', whereby autonomous robots learn to modify their behaviour by trial-and-error, is a significant area of research. L. Steels, Ten Big Ideas of Artificial Intelligence, Remarks to the 25th Benelux Conference on Artificial Intelligence, Delft Technical University, 8 November 2013; author interview with G. Visentin, Head of the Automation and Robotics Department, European Space Agency, Noordwijk, 4 November 2013; P. Margulies, Making Autonomous Weapons Accountable: Command Responsibility for Computer-Guided Lethal Force in Armed Conflict, in: J. Ohlin (ed.) Research Handbook on Remote Warfare, Cheltenham 2016 (forthcoming), pp. 7-12.

15 R. Brooks, A Brave New World? How Will Advances in Artificial Intelligence, Smart Sensors and Social Technology Change Our Lives?, Panel Discussion at the World Economic Forum, 22 January 2015, https://www.youtube.com/watch?v=wGLJXO08Iyo.

16 'Tactical patience' refers to the ability to permit a combat situation to develop to ensure that actions taken (such as attacks) are appropriate and lawful. See T. McHale, Executive Summary for AR 15-6 Investigation, 21 February 2010 CIVCAS Incident in Uruzgan Province, Memorandum for Commander, United States Forces-Afghanistan/International Security Assistance Force, Afghanistan, http://www.rs.nato.int/images/stories/File/April2010-Dari/May2010Revised/Uruzgan%20investigation%20findings.pdf.

force, when engaging the enemy, and so put civilians at lesser risk.[17] Consequently, the use of these autonomous systems will, in some situations, impact the judgment process of balancing military necessity[18] and humanity[19] embodied in proportionality analysis.[20] Indeed, the introduction of these weapons to the battlespace can alter the meaning and scope of these two principles.[21]

In parallel, the employment of autonomous weapon systems can change the scope and nature of judgment required to comply with human rights law. Although international humanitarian law and international human rights law are distinct bodies of law, they protect similar principles and interests[22] and, in a normative sense, modern international humanitarian law has roots in international human rights law.[23] Thus, "(...) humanitarian law also

17 M. Sassòli (n 10), p. 310; M. Schmitt and J. Thurnher (n 10), p. 264.

18 Francis Lieber defined 'military necessity' as "the necessity of those measures which are indispensable for securing the ends of the war, and which are lawful according to the modern law and usages of war"; see General Orders No. 100, Instructions for the Government of Armies of the United States in the Field, http://avalon.law.yale.edu/19th_century/lieber.asp#art1. The UK armed forces use a more nuanced definition: "Military necessity permits a state engaged in an armed conflict to use only that degree and kind of force, not otherwise prohibited by the law of armed conflict, that is required in order to achieve the legitimate purpose of the conflict, namely the complete or partial submission of the enemy at the earliest possible moment with the minimum expenditure of life and resources." See JSP 383, The Joint Service Manual of the Law of Armed Conflict (2004 ed.), Joint Doctrine and Training Centre, UK Ministry of Defence, para. 2.2, https://www.gov.uk/government/publications/jsp-383-the-joint-service-manual-of-the-law-of-armed-conflict-2004-edition.

19 The principle of 'humanity' prohibits the infliction of suffering, injury or destruction not actually necessary for the accomplishment of a legitimate military purpose. See ibid, paras. 2.4 and 2.4.1.

20 May and Newton suggest that the time has arrived to consider, as lex ferenda, the lives of combatants as factors in a proportionality assessment. See L. May and M. Newton, Proportionality in International Law, New York 2014, p. 151. In that context, in certain circumstances, particularly when capture is possible, there may be little military advantage to be gained from the use of lethal force by autonomous weapon systems against, or in the vicinity of, human soldiers.

21 The notions of military necessity and humanity can evolve as new technology affects the ways wars can be fought and social perceptions of acceptable human suffering change. See H. Natsu, Nanotechnology and the Future of the Law of Weaponry, 91 International Law Studies 486 (2015) 501-502, 507.

22 See ICTY, Judgment, Prosecutor v. Zejnil Delalić, et. al, IT-96-21-A, 20 February 2001, para. 149: "Both bodies of law take as their starting point the concern for human dignity, which forms the basis of a list of fundamental minimum standards of humanity."

23 See ICTY, Decision on the Defence Motion for Interlocutory Appeal on Jurisdiction, Prosecutor v Dusko Tadic, IT-94-1, 2 October 1995, para. 87: the idea of 'international humanitarian law' "has emerged as a result of the influence of human rights doctrines on the law of armed conflict". An important feature of both 1977 Additional Protocols "is the way in which their content has been influenced by the law relating to human rights". See Prefatory Note to Chapter 24, 1977 Geneva Protocol I Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, in: A. Roberts and R. Guelff (ed.), Documents on the Laws of War, 3rd ed., Oxford 2000, p. 419.

contains a prominent human rights component",[24] and human rights law continues to apply during armed conflict.[25] The precise contours, however, of the application of international human rights law during combat, and its interplay with international humanitarian law, remain a matter of debate.[26] The European Court of Human Rights holds, rather vaguely, that during armed conflict, relevant provisions of international human rights covenants "should be accommodated, as far as possible" with the relevant laws of war.[27]

Thus, in the context of armed conflict, 'judgment' requires the ability to integrate and apply at least two different bodies of law. Unlike international humanitarian law, which permits soldiers to kill their enemy unless they are hors de combat, international human rights law permits the exercise of lethal force only when 'absolutely necessary'.[28] Consistent with this human rights standard, the International Committee of the Red Cross (ICRC) suggests that during armed conflict, international law prohibits (or should prohibit) soldiers from killing enemy combatants when the possibility of capture or other non-lethal means to neutralise the enemy exists.[29]

24 Ibid, p. 10.

25 ICJ, Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory, Advisory Opinion, 9 July 2004, para. 106. The Court held that Israel, by constructing a wall that passed through occupied territory, breached various obligations under the applicable international humanitarian law and international human rights law, see para. 137; see also ICJ, Democratic Republic of the Congo v. Uganda, 19 December 2005, paras. 179 and 216–220; ICJ, Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 8 July 1996, para. 25; ECtHR, Case of Al-Jedda v. The United Kingdom, Judgment, No. 27021/08, 7 July 2011, para. 105.

26 See e.g. ECtHR, Case of Hassan v. The United Kingdom, Judgment, No. 29750/09, 16 September 2014, paras. 101–107.

27 Ibid, para. 104; also see EWHC, Serdar Mohammed v. Ministry of Defence, Judgment, [2014] EWHC 1369 (QB), 2 May 2014, para. 288, holding that in a situation where a more specialised body of international law also applies, provisions of the Convention should be interpreted as far as possible in a manner consistent with that lex specialis.

28 ECtHR, Case of McCann and Others v. The United Kingdom, Judgment, No. 18984/91, 27 September 1995, para. 214; see Art. 2 ECHR.

29 N. Melzer, Interpretive Guidance on the Notion of Direct Participation in Hostilities in International Humanitarian Law, May 2009, pp. 77, 82. The Israeli Supreme Court has taken a similar position with respect to the possible arrest of suspected terrorists, in particular under conditions of belligerent occupation. Thus, in Israeli domestic law, "among the military means, one must choose the means whose harm to the human rights of the harmed person is smallest", HCJ, The Public Committee Against Torture in Israel v. The Government of Israel ('Targeted Killing Case'), Judgment, HCJ 769/02, 11 December 2005, para. 40. Importantly, the Court based its decision on "the rules of international law" as well as Israeli law, see para. 61.

Time will tell whether states will adopt the Interpretive Guidance's more restrictive approach to the use of lethal force during armed conflict.[30] The increasing use of autonomous weapon systems, however, could facilitate the development of customary law in this direction. The availability of autonomous weapon systems with the capacity to make these value judgments puts fewer human soldiers at risk from enemy combatants and reduces the dangers of efforts to capture the enemy (i.e. the dangers of accommodating international human rights law).[31] Thus, by fielding sophisticated autonomous weapon systems, belligerent parties reduce the number of situations where the use of lethal force is 'absolutely necessary'. Modern armed forces that follow the ICRC's standard and field lethal autonomous weapon systems will constantly assess (using human judgment, artificial intelligence, or both) whether it is inappropriate (if not illegal) to use lethal force instead of capturing enemy combatants.

It remains to be seen whether artificial intelligence software can advance to a level whereby computers can make complex, value-based determinations as to the 'absolute necessity' of using lethal force against enemy targets to serve 'a legitimate military purpose' in the 'prevailing circumstances'. Should the Interpretive Guidance's standard for the use of lethal force one day become an obligation under customary international humanitarian law, the fielding of lethal autonomous weapon systems lacking this capacity would violate the law's prohibition of means and methods of warfare designed to cause unnecessary suffering.[32] Under the same standard, use of such weapons would violate international human rights law's proscription of arbitrary deprivations of the right to life.

30 Part IX of the ICRC's Interpretive Guidance has been the subject of sustained and forceful criticism, in particular from international humanitarian law experts with military expertise, who contend that the ICRC incorrectly imposes international human rights standards onto the norms and obligations of the law of war; see e.g. M.N. Schmitt, The Interpretive Guidance on the Notion of Direct Participation in Hostilities: A Critical Analysis, 1 Harvard National Security Journal 5 (2010); W.H. Parks, Part IX of the ICRC "Direct Participation in Hostilities" Study: No Mandate, No Expertise, and Legally Incorrect, 42 International Law and Politics 769 (2010); K. Watkin, Opportunity Lost: Organized Armed Groups and the ICRC "Direct Participation in Hostilities" Interpretive Guidance, 42 International Law and Politics 641 (2010); for a more positive view see the remarks of R. Goodman, The Changing Character of the Participants in War: Civilianization of War-Fighting and the Concept of "Direct Participation in Hostilities", US Naval War College, International Law Conference 2010, https://archive.org/stream/internationallaw87pedr/internationallaw87pedr_djvu.txt.

31 For a discussion of the related issue of the capacity of autonomous weapon systems to recognise the intent of combatants to surrender, see R. Sparrow, Twenty Seconds to Comply: Autonomous Weapon Systems and the Recognition of Surrender, 91 International Law Studies 699 (2015).

32 Article 35(2) API.

Thus, the introduction of autonomous weapon systems into the battlespace changes the scope and nuances of the judgments required to be made by the persons who field them. Under this perspective of judgment and autonomy, it is short-sighted to suggest that the importance and input of human judgment can be minimised in the design and fielding of autonomous weapon systems.[33] Where complex (and sometimes conflicting) values are at stake, priority should be given to the reinforcement of human-machine teamwork rather than separation of duties between humans and machines.[34] Both human judgment, and autonomous technologies, are tools. Autonomy, therefore, can be viewed not as an end in itself, but as a tool to accomplish particular objectives,[35] with the input and assistance of human judgment when necessary.

However, even the most sophisticated and "flawless technology of man"[36] can produce unforeseen injury to humankind. In the case of lethal autonomous weapon systems, we can *perceive* the damage that will be done to the evolution of human judgment by the use of these weapons. The increasing speed of communications, data processing, and autonomous weapon technology shortens the time available for manned and unmanned weapon systems to react to events and, when necessary, attack enemy combatants and objectives. The inevitable velocity of autonomous military engagements will obstruct the development (as well as the use) of sound human judgment that arises from opportunities for human reflection on one's own important experiences and those of others.[37] The foreseeable danger is that, as the velocity of armed conflict escalates, human judgment will lose its relevance.

33 J. Bradshaw et al., The Seven Deadly Myths of "Autonomous Systems", Human-Centred Computing, May/June 2013, p. 57, www.jeffreymbradshaw.net/publications/IS-28-03-HCC_1.pdf.

34 Ibid, pp. 58-60; as a team, humans and computers are far more powerful than either alone, especially under uncertainty. See M. Cummings (n 2), p. 12. For example, if autonomous weapon systems can exercise 'self-recognition', i.e. the capacity to detect that it is operating outside the conditions for which it was designed, the machine will call on humans for increased supervision; author interview with M Johnson (n 1).

35 Author Interview with G. Visentin (n 14).

36 P. Mahon, Royal Commission of Inquiry into and Report Upon the Crash on Mount Erebus, Antarctica, of a DC10 Aircraft Operated by Air New Zealand Limited, Wellington 1981, para. 398, http://www.erebus.co.nz/LinkClick.aspx?fileticket=PUWvCWDoUoE%3D&tabid=159.

37 Colonel Shane Riza, a U.S. Air Force fighter pilot, explains that presently military "communication occurs at the speed of light" and recognises that autonomous weapons systems permit "the speed of future decision cycles outpacing the human mind", see M. Shane Riza, Killing Without Heart: Limits on Robotic Warfare in an Age of Persistent Conflict, Washington 2013, p. 41. In the future, for example, operations of the Israeli Defence Forces will include swarms of autonomous land, air and sea weapon systems, including networks of miniature and nanotechnology platforms, see Lt. General Gantz, The IDF in 2025, Address to the Begin-Sadat Centre for Strategic Studies, 9 October 2013, https://www.youtube.com/watch?v=vPAEkk5LOXc.

Autonomous Vehicle Systems in the Civil Sphere

David Shim*

Introduction

UAVs first found their use in military applications just like many other modern technologies such as computers or antibiotics. They were initially developed for aerial surveillance and proved very useful during the Kosovo conflict in the late 1990s, as well as the wars in Afghanistan and Iraq in the early 2000s. Subsequently, some people started thinking that UAVs could be also employed as civil applications, for instance border control or long-endurance surveillance.

An unmanned aerial vehicle consists of a flight control computer, navigation sensors, and communication devices to enable automatic flight while communicating with the human operator(s) at a remote place. This concept of operation, and the vehicle's key enabling technologies, are identical in both military and civil applications. Thus, the technology is inherently dual-use. One key difference is that UAVs for civil application need to co-operate with other existing manned aircraft in the civil airspace. In addition, due to their unproven reliability and vulnerability to cyber-attacks and hacking, there have been considerable efforts to provide various rules and standards for full-scale deployment. Many countries have started their own effort, and recently they have joined under the leadership of the International Civil Aviation Organization (ICAO). Within this framework, only UAVs capable of complying with existing rules and standards for full-scale manned aircraft are allowed to be integrated into the civil airspace. The making of specific rules is currently underway, and it is anticipated that those new international rules will be ready by 2020.

Recently, as an offspring from the lineage mentioned above, a different kind of UAVs have emerged as a relevant player. Most easily recognised by multiple counter-rotating rotors, these UAVs are often known as 'drones'. These drones utilise very small but powerful sensors and CPUs originally developed for smartphones. Therefore, they are much cheaper yet very easy to operate, so they are now widely used for aerial photography. Amazon, the world's largest internet retailer, even seeks to use these drones for aerial delivery. By using such devices, they could revolutionise the logistics process, greatly shorten the delivery

* KAIST, Daejeon, South Korea.

time, and lower the shipping cost. These small drones are also very popular among hobbyists, and are used for a number of applications that could not be handled by large manned aircraft. However, these drones are not very reliable, and moreover could be used by people with malicious intentions – just like their larger cousins, but even more easily. Therefore, very common uses of these drones beyond visual range are prohibited by authorities until safety can be guaranteed by better technologies and enforceable rules, which are not fully developed yet. These multirotor drones were developed separately from military applications, but some of them are finding their way into military applications. Adequately modified, these drones are expected to perform close-range surveillance and even ordnance delivery for military employment with far greater ease and much lower cost.

At this moment, an explosion of UAVs for civil applications is just about to happen. Global collaboration on the making of rules and providing infrastructures for UAVs is actively ongoing. In this context, it is crucial to note that, under the concept of operation pursuant to the ICAO, these UAVs are not allowed to operate 'autonomously'. Currently, these UAVs are integrated into the existing airspace on the condition that they oblige to existing 'rules of air', as specified in the ICAO Conventions and Annexes. Within this framework, UAVs are not allowed to operate autonomously because, if they did, air traffic controllers could not predict how they will behave in relation to other nearby aircraft, and would therefore fail to guarantee safety.

However, this does not mean these UAVs cannot operate autonomously at all. Recently, advances of small but powerful sensors and computers equipped with the latest technologies such as deep learning, these UAVs *could* perform autonomous tasks if permitted accordingly. Small UAVs (drones) are also needed to be used in a segregated airspace (as they cannot meet ICAO standards). These drones cannot meet all the performance requirements to enter the civil airspace: they are small – so they may not be visible to other aircraft – far less reliable, and lack in the communication and other safety equipment due to their limited payload. Therefore, these drones are currently limited to operate only in the visible range and typically in an altitude of less than 150 metres during daytime. However, if the reliability is being improved in the future and can be properly monitored, those qualified drones could be allowed to operate beyond visual range or even at night-time. Once that happens, these small drones will find new modes of employment. However, when they are deployed on a large scale, it will not be easy to control them in the conventional air traffic management. For this reason, some researchers propose the development of an autonomous air traffic control system that is driven by a highly sophisticated computer program which communicates with the participating UAVs with high-speed wireless communication such as 4G LTE or beyond. This way, they hope that UAVs can operate in a much safer

and efficient way for full-scale deployment. In this construct, they are allowed to operate as autonomously as possible, since it is not so easy to maintain tight communication with overwhelming number of UAVs concurrently operating within the same area.

As drones are designed to follow the orders from remote operators, they almost always have open communication channels. Normally, this access channel should be guarded by proper security protocols. For now, as drones are at their infancy, actual security levels are not very high. Soon, however, it will be required to apply proper security measures to their communication hardware and software. Security is a very serious concern for larger UAVs which can cause much bigger damage. In addition to cybersecurity, which protects UAVs from malicious external attacks, some users may decide to use them for criminal activities such as transporting illegal goods across borders, lethal weapons, or even small bombs. And although rule-makers will certainly try very hard to propose appropriate laws and security guidelines, it is to be feared that completely stopping criminal activities using drones will be virtually impossible.

Self-driving cars

Self-driving cars are another hot topic in the field of autonomous vehicles. The desire for such cars has been around for a very long time, but the autonomous driving as we know it today became a reality with the DARPA 'Grand Challenge' and 'Urban Challenge' (2004-2007). DARPA, a U.S. defence research agency, was seeking to develop (or promote) autonomous driving technology because many soldiers were lost during ground transportation, either by improvised explosive devices or accidents. Prior to this event, it was common belief that a car cannot be driven autonomously without some external aids, such as roads with especially embedded sensors, and communications from other cars. However, similarly to the case with UAVs, thanks to the advances in sensor and computer technology, the competing cars could run for 240 kilometres without any intervention in the 2005 competition. Later, in 2007, during the DARPA 'Urban Challenge' the cars were required to drive in an urban area for 96 kilometres while obeying Californian road traffic regulations. Though it had been questioned whether it would even be possible to meet the requirements, four teams were able to finish the challenge. The respective technology was transferred to many industries, most notably Google, and today it is widely believed that autonomous cars are on the verge of becoming a reality for everyone. Some of these technologies are becoming available as ADAS (Advanced Driver Assistance System), such as adaptive cruise control or active lane-keeping. In 2015, Tesla, an electric car manufacturer, introduced 'Autopilot', which allows the car to drive on an urban freeway with unprecedented level of autonomy.

Autonomous cars use various sensors to collect *in-situ* information on the surrounding environment including roads and other cars, and find a safe path to the destination while obeying various traffic rules. Cameras, laser scanners, radars, and ultrasonic sensors can be used in combination in order to build a map of the surroundings. Communication with other cars, nearby infrastructure, and cloud servers for the maps and traffic conditions can be of great help to lower risk and to improve efficiency. Since there is no driver for fully autonomous cars, the driving system must be autonomous. Many cutting-edge technologies such as GPU-aided computation and deep learning will help to improve accuracy and safety.

For now, although leading companies such as Google assert that autonomous cars will arrive in the near future, due to the complexity and uncertainty of the real world – further challenged by legal and ethical issues – fully autonomous cars will not be readily available anytime soon. However, cars with quite high levels of autonomy will become a reality, possibly much sooner than many people expect. Drivers will get benefits such as lower work load and higher safety.

In 2016, Uber and others have been testing their versions of autonomous cars on public roads, offering test drives. They are currently monitored by engineers who take over control in case of an emergency. The test cars are very expensive due to many high-quality sensors and ongoing technology development. Still, once they become popular and if technology continues to advance, they are expected to be quite affordable.

Many people raise questions concerning the safety of self-driving cars. The simple answer is that they will not be used until they have become safer than typical human-driven cars. Also, some people challenge whether the autonomous driving algorithms can make rational decisions on how to 'minimise' the damage when the car has no choice but to hit a person in order to save the occupants of the car, or the other way around. This might be, however, the wrong question to ask: cars have always been developed to protect their occupants, and others should do the same to protect themselves, including the pedestrians. Moreover, autonomous cars may not even enter such a situation because they always abide by the traffic rules and speed limits, unlike human drivers, who speed or otherwise ignore the rules. What is more, autonomous cars have much longer detection ranges, much faster decision-making capabilities, and also much more precise control. Therefore, the overall accident rate will be greatly lowered, raising questions whether humans should still be allowed to drive cars.

Autonomous cars are also expected to lower demand, since they can be shared by multiple family members. Large parking lots will not be needed anymore, so the city design will be revolutionised and pollution will be lowered while the existing automakers will suffer from greatly diminished demands.

Autonomous cars are dual-use. The technology we know today was explored for military applications. Although the technology has matured more into civil sectors, it could find its way back to military applications (such as convoy trucks). It could be also adapted to be used for autonomous tanks and armoured vehicles, and even for self-detonating ground weapons.

Comparison between UAVs and self-driving cars

The biggest difference between UAVs and self-driving cars is that, in terms of maturity, while UAVs already have the necessary technology but need to have it perfected for safety reasons, self-driving cars do not yet have the technology for fully autonomous operation. The conditions on the road are much more complex and unpredictable than those in the air. At the same time, aviation accidents can be very damaging to many people. However, one can be carefully optimistic for the future of UAVs thanks to the active development of international rules and regulations. Self-driving cars, on the other hand, will need some more time to become a full reality, aided by more powerful computers, sensors, and advanced algorithms.

Autonomous operation of UAVs will not be allowed in the near future, but as the technology matures, someday soon, even 'manned drones' may emerge. Those are drones that transport human passengers, who do not (cannot) pilot the vehicle themselves. Such drones can be flown remotely by a pilot, or even fully autonomously quite similar to autonomous cars. These two areas will help each other develop, leading to fully autonomous vehicles in a not so distant future.

Situational Awareness and Adherence to the Principle of Distinction as a Necessary Condition for Lawful Autonomy

Lucy Suchman*

Introduction

The questions surrounding lethal autonomous weapons systems (LAWS) are being addressed by the Convention on Certain Conventional Weapons (CCW) along multiple lines of analysis. This chapter is meant as a contribution to discussions regarding the concept of 'autonomy', on the basis of which I present an argument questioning the feasibility of LAWS that would comply with International Humanitarian Law (IHL).[1] This argument is based not on principle, but rather on empirical evidence regarding the interpretive capacities that legal frameworks like IHL presuppose for their application in a specific situation. These capacities make up what in military terms is named 'situational awareness'.[2] Despite other areas of progress in artificial intelligence (AI) and robotics, it is my assessment that none has been made in the operationalisation of situational awareness in an indeterminate environment of action. More specifically for the question of LAWS, situational awareness as a prerequisite for the identification and selection of legitimate targets – what has been named the 'principle of distinction' – is not translatable into machine-executable code. Yet situational awareness is essential for adherence to IHL or any other form of legally accountable rules of conduct in armed conflict.

* Professor of Anthropology of Science and Technology, Lancaster University, UK.

1 International humanitarian law "is a set of rules which seek, for humanitarian reasons, to limit the effects of armed conflict. It protects persons who are not or are no longer participating in the hostilities and restricts the means and methods of warfare". See https://www.icrc.org/eng/assets/files/other/what_is_ihl.pdf.

2 Situational awareness can be defined as "understanding of the operational environment in all of its dimensions – political, cultural, economic, demographic, as well as military factors"; B.C. Dostal, Enhancing Situational Understanding Through the Employment of Unmanned Aerial Vehicles, Center for Army Lessons Learned, 2001, http://www.globalsecurity.org/military/library/report/call/call_01-18_ch6.htm.

This assessment is based on my position as an anthropologist engaged for over three decades with the fields of AI and human-machine interaction.[3] A central aim of my remarks (and of my larger body of research) is demystification of the field of AI, particularly with respect to clarification of the differences between human and machine capabilities. My work in tracking developments in AI and robotics involves taking seriously the claims that are made for intelligent machines and comparing them to extensive studies of the competencies – perceptual, and also crucially social and interactional – that are the basis for associated human activities. My focus on situational awareness in the context of this publication arises not only from the fact that it is a prerequisite for lawful action within the framework of IHL, but also because this is an area in which I hope that my particular perspective can contribute to greater clarity on the key concept of autonomy.

LAWS and the principle of distinction

The elements of situational awareness that I believe are most relevant to the question of whether LAWS can be adherent to IHL are those that inform the requirement of distinction in the use of lethal force; that is, discrimination between legitimate and non-legitimate targets.[4] I recognise that the requirements of distinction and proportionality are closely linked, but insofar as proportionality judgments presuppose that distinction has been made, I focus on distinction here. In the case of autonomous weapons, adherence to the principle of distinction would require that robots have adequate vision or other sensory processing systems, and associated algorithms, for separating combatants from civilians and for reliably differentiating wounded or surrendering combatants from those who are not. Existing machine sensors such as image-processing cameras, infrared temperature sensors, and the like, may be able to identify something as a human, but they cannot make

3 Before taking up my present Chair at Lancaster University, I was a Principal Scientist at Xerox's Palo Alto Research Center, where I spent twenty years as a researcher. I have written for both social and information sciences audiences, including two books, Human-Machine Reconfigurations, Cambridge 2007, and Plans and Situated Actions: The Problem of Human-Machine Communication, Cambridge 1987. In 2002, I received the Benjamin Franklin Medal in Computer and Cognitive Sciences, and in 2010 the ACM SIGCHI Lifetime Research Award. In 1983, I was a founding member of Computer Professionals for Social Responsibility, an organisation formed to address the increasing reliance on computing in the control of nuclear weapon systems; I am now a member of the International Committee for Robot Arms Control, and on the Executive Board of the Foundation for Responsible Robotics.

4 On the Principle of Distinction see http://www.icrc.org/customaryihl/eng/docs/v1_cha_chapter1_rule1. Also see D. Garcia, Future arms, Technologies, and International Law: Preventive Security Governance, 1 European Journal of International Security 94 (2016) 96, in which she calls for what she terms 'preventive security governance', focused on the three areas of (1) preventing future harm to civilians, (2) responsibility and accountability, and (3) the question of what constitutes a legal and legitimate attack.

the discriminations among persons that are required by the principle of distinction.[5] Even if machines had adequate sensing mechanisms to detect the difference between civilians and uniform-wearing military, they would fail under situations of contemporary warfare where combatants are most often not in uniform.[6] And more sophisticated technologies, such as facial or gait recognition, are still reliant on the existence either of a pre-established database of templates, against which a match can be run, or profiles, which are inherently vulnerable to false positives and other forms of inaccurate categorisation.[7]

I take as a working definition of LAWS weapons systems in which the identification and selection of human targets and the initiation of violent force is carried out under machine control; that is, these capacities are delegated to the system in ways that preclude deliberative and accountable human intervention or what, in the current discussion, has been

5 Some opponents of a ban on LAWS imagine scenarios in which the mere presence of a human body is an adequate criterion for the identification of that person as a legitimate target. But that requirement is counter to the direction in which conflict is moving, as the boundaries that designate geographic zones of combat are increasingly fluid and contested.

6 On the increasing complexity of the combatant/civilian distinction, Wilke observes that "the rise of the figure of the 'unlawful combatant' (...) is accompanied by a corresponding rise of the figure of the illegitimate, noninnocent, suspicious civilian"; C. Wilke, Civilians, Combatants and Histories of International Law, Critical Legal Thinking, 28 July 2014, http://criticallegalthinking.com/2014/07/28/civilians-combatants-histories-international-law/.

7 With respect to the development of algorithmic templates for the identification of legitimate targets, Schuppli observes that "the recently terminated practice of 'signature strikes' in which data-analytics were used to determine emblematic 'terrorist' behaviour and match these patterns to potential targets on the ground already points to a future in which intelligence gathering, assessment, and military action, including the calculation of who can legally be killed, will largely be performed by machines based upon an ever expanding database of aggregated information"; S. Schuppli, Deadly Algorithms, Continent, Issue 4.4, 2015, pp. 20-27, available at http://www.continentcontinent.cc/index.php/continent/article/view/212. The concern here is with an increasing push towards reliance on *a priori* stereotyping, rather than systematic intelligence gathering; it is the unreliability of stereotyping that has discredited this practice. On the exacerbation of the problem of targeted killing by LAWS see also C. Heyns, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, UN Doc. A/HRC/23/47, 2013.

characterised as 'meaningful human control'.[8] This definition follows the one adopted by UN Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns, who defines LAWS as "robotic weapons systems that once activated, can select and engage targets without further intervention by a human operator".[9] The emphasis in this discussion is specifically on *human* targets; that is, the identification of humans or human-inhabited objects (buildings, vehicles) as lawful targets for engagement. I am excluding, in other words, defensive weapon systems that operate on the basis of unambiguous signals from another (unmanned or uninhabited) device that comprises an imminent threat.

The fundamental problem in meeting the requirements of the principle of distinction is that we do not have a definition of a civilian that can be translated into a recognition algorithm. Nor can we get one from IHL.[10] The 1949 Geneva Conventions require the use of 'common sense,' while the 1977 Protocol I essentially defines a civilian in the negative

8 For the minimum necessary conditions for meaningful human control see Article 36, Killer Robots: UK Government Policy on Fully Autonomous Weapons, 2013, http://www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf; Article 36, Key Areas for Debate on Autonomous Weapon Systems: Memorandum for Delegates at the Convention on Certain Conventional Weapons, paper presented at the Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, 13-16 May 2014, www.article36.org/wp-content/uploads/2014/05/A36-CCW-May-2014.pdf; F. Sauer, ICRAC Statement on Technical Issues to the 2014 UN CCW Expert Meeting, ICRAC, 14 May 2014, http://icrac.net/2014/05/icrac-statement-on-technical-issues-to-the-un-ccw-expert-meeting. The word 'meaningful' here is meant to anticipate and reject the proposition that any form of oversight over automated target identification constitutes adequate human control. Horowitz and Scharre propose that in its emerging usage, "meaningful human control has three essential components: human operators are making informed, conscious decisions about the use of weapons; human operators have sufficient information to ensure the lawfulness of the action they are taking, given what they know about the target, the weapon, and the context for action; and the weapon is designed and tested, and human operators are properly trained, to ensure effective control over the use of the weapon"; M. Horowitz and P. Scharre, Meaningful Human Control in Weapon Systems: A Primer, Center for a New American Security Working Paper, March, 2015, https://s3.amazonaws.com/files.cnas.org/documents/Ethical_Autonomy_Working_Paper_031315.pdf, p. 4.

9 Heyns (n 7), para. 133; see also Sharkey, Automating Warfare: Lessons Learned from the Drones, 21 Journal of Law, Information and Science 2012; Scharre and Horowitz (n 8), p. 16, offer a closely related definition, but one focused more specifically on the question of targeting, viz.: an "autonomous weapon system is a weapon system that, once activated, is intended to select and engage targets where a human has not decided those specific targets are to be engaged". Addressing the key phrase 'select and engage', Gubrud observes that 'selection' or targeting is complicated by the fact that "the status of an object as the target of a weapon is an attribute of the weapon system or persons controlling and commanding it, not of the object itself". Target selection, Gubrud argues, is where the crucial questions and indeterminacies lie, and the operator, "the final human in the so called 'kill chain' or 'loop'", should be the final decision point; M. Gubrud, Autonomy Without Mystery: Where Do You Draw the Line?, 9 May 2014, http://gubrud.net/?p=272.

10 See P. Asaro, How Just Could a Robot War Be?, in P. Brey et al. (ed.), Current Issues in Computing And Philosophy, Amsterdam 2009, pp. 50-64. He reminds us that IHL comprises a diverse body of international laws and agreements (such as the Geneva Conventions), treaties, and domestic laws. These are far from algorithmic specifications for decision-making and action.

sense, as someone who is not a combatant.[11] While robotics may achieve effective sensory and visual discrimination in certain narrowly constrained circumstances, human level discrimination with adequate common sense reasoning for situational awareness would appear to be computationally intractable.[12] At this point, at least, there is no evidence or research result to suggest otherwise.

Relations of automation and autonomy

While drawing a line between automation and autonomy is necessary in the context of the CCW's deliberations, this does not imply that autonomous systems are not automated. The crucial question, rather, is whether or not an automated system is subject to meaningful human control.[13] We could, in other words, define autonomous systems precisely as those in which the identification, selection and engagement of targets has been fully automated – this definition still provides a clear distinction between automated systems under human control and those that are not (i.e. weapons systems that are acting autonomously).

It is also the case that the question of autonomy with respect to LAWS needs to be considered within a longer history of the intensifying automation of warfare. This is a trajectory justified as a necessary response to the demand for increasingly rapid engagement, along with the vulnerabilities incurred through reliance on complex information and communications networks; a problem that greater automation and system complexity further exacerbates.

We have seen these dynamics before in the case of launch on warning in nuclear weapon systems, and some of the questions currently under debate were addressed, and arguably resolved, in the work of computer scientists like David Parnas in the 1980s.[14] In the context of the U.S. Strategic Defense Initiative, Parnas made the crucial distinction between a computational system's verifiable execution of its specifications on one hand (what is commonly referred to as the software's 'correctness', or reliability in the narrow sense described by

11 Art. 50(1) of the Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts, 8 June 1977.

12 See N. Sharkey, Grounds for Discrimination: Autonomous Robot Weapons, 11 RUSI Defence Systems 86 (2008).

13 In a distinction consistent with the definition adopted here, Scharre and Horowitz (n 8), p. 17, propose the crucial difference as that between a weapon that is selecting targets without human decision ('self-targeting' or autonomous) and a weapon engaging a human-selected target. See also Gubrud (n 9).

14 See D.L. Parnas, Software Aspects of Strategic Defense Systems, 28 Comm ACM 1326 (1985); see also B.C. Smith (December 1984), The Limits of Correctness., 14/15 ACM SIGCAS, Computers and Society (1985).

Asaro),[15] and the system's ability to assess the conditions in which those specifications apply on the other (necessary for its reliability in operation). Simulated testing of automated weapons systems can assess correctness, but it can never definitively assure reliability under actual conditions of use. The only way to achieve the latter is through practical methods of iterative development based on repeated trials under conditions that closely match those of intended deployment, or informed by experience of the system in use, neither of which is possible in the case of complex weapons systems with deadly consequences. It was for this reason, among others, that the Strategic Defense Initiative was finally abandoned.

The U.S. Department of Defense's 'Unmanned Systems Integrated Roadmap 2011-2036' distinguishes automatic from autonomous systems in this passage:

> *"Dramatic progress in supporting technologies suggests that unprecedented levels of autonomy can be introduced into current and future unmanned systems (...). Automatic systems are fully preprogrammed and act repeatedly and independently of external influence or control (...). However, the automatic system is not able to define the path according to some given goal or to choose the goal dictating its path. By contrast, autonomous systems are self-directed toward a goal in that they do not require outside control,* but rather are governed by laws and strategies that direct their behaviour (...). The special feature of an autonomous system is its ability to be goal-directed in unpredictable situations*. This ability is a significant improvement in capability compared to the capabilities of automatic systems."*[16]

The key phrase here is "governed by laws and strategies that direct their behaviour (...) in unpredictable situations". As I have stated above, 'laws and strategies' are not translatable to executable code. In assessing the feasibility of the system posited in this passage, it is crucial to keep in mind that autonomy or 'self-direction' in the case of machines presupposes the unambiguous specification (by human designers) of the conditions under which associated actions should be taken. And this requirement for unambiguous specification of condition/action rules marks a crucial difference between autonomy as a human capacity, and machine autonomy. As I have argued in previous writing, autonomy as we understand it in the context of human action means self-direction under conditions that are not, and cannot be,

15 P. Asaro, Roberto Cordeschi on Cybernetics and Autonomous Weapons: Reflections and Responses, Paradigmi: Rivista di critica filosofica, Anno XXXIII, no. 3, Settembre-Dicembre 2015, pp. 83-107, 90.

16 U.S. Department of Defense, Unmanned Systems Integrated Roadmap FY2011-2036, p. 43 (emphasis added).

fully specified by rule.[17] This in turn accounts for what we might call the 'strategic vagueness' of any kind of rule or directive for action; that is, the assumption that the exercise of the rule, or the execution of the directive or plan, will involve *in situ* judgment regarding the rule's application. Where the requisite competencies are in place, this openness – far from being a problem – is what enables the effectiveness of a general plan or rule as a referent for situated action.

Limits to information processing as a model of situational awareness

To make this more concrete, we can take the case of the human 'action according to rules', which defines military discipline, and is most pertinent to this discussion of IHL and the rule of distinction. Because the precise conditions of action in combat cannot be known in advance, the directives for action in the case of military operations presuppose competencies for their accurate 'execution' that the directive as such does not and cannot fully specify. Thus the requirement for situational awareness as necessary to effective and, most importantly for our purposes, legally accountable warfare.

Approaches to AI-based robotics share the common requirement that a machine can engage in a sequence of 'sense, think, and act'.[18] It is crucial in this context, however, to be wary of the use of evocative terms that imply the functionality of programs, rather than providing technical descriptions of actual capabilities.[19] Does 'sense, think, and act' refer to an assembly line robot that performs an action at a fixed location, in relation to an environment carefully engineered to match its sensing capacities, and where the consequences of failure are non-lethal? Or does it invoke sensing and perception as dynamic and contingent capacities, in open-ended fields of (inter)action, with potentially lethal consequences? In the case of human combatants, the ability to be goal-directed in unpredictable situations presupposes capacities of situational assessment and judgment, in circumstances where the range of those capacities is necessarily open ended. Combat situations, moreover,

17 See L. Suchman, Plans and Situated Actions: The Problem of Human-Machine Communication, Cambridge 1987; and L. Suchman, Human-Machine Reconfigurations, Cambridge 1987. This problem is not resolved by the promises of 'machine learning' to enable derivations from data external to the specified rule, insofar as learning algorithms continue to rely upon the availability of specified data formats rather than open-ended horizons of input, see Asaro (n 15).

18 See L. Suchman and J. Weber, Human-Machine Autonomies, in N. Bhuta et al. (ed.), Autonomous Weapon Systems: Law, Ethics, Policy, Cambridge 2016, pp. 75-102.

19 See N. Sharkey and L. Suchman, Wishful Mnemonics and Autonomous Killing Machines, 136 AISBQ Quarterly, the Newsletter of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour 14 (2013).

frequently involve opponents who work hard and ingeniously to identify and defeat any prior assumptions about how they will behave. This is in marked contrast to the situations in which AI techniques, and automation more generally, have been successfully applied. In any case, the burden of proof here must rest with proponents, and require a higher standard than general assertions of progress in artificial intelligence, which is debatable other than in certain, limited technical areas that do not yet begin to address problems of reliable discrimination between legitimate and illegitimate human targets.[20]

A final note is that autonomy is best understood not as an individual capacity – whether human or machine – but rather as a capacity enabled by particular configurations of people and technologies. Different configurations make different capacities for action possible. In thinking about life-critical technical systems, the key question is what conditions a particular configuration affords for human responsibility and accountability. This is where the concept of meaningful human control becomes crucial: what is required to ensure that delegations of agency to machines allow the preservation of human responsibility and accountability? In a report issued in February of this year,[21] UN Special Rapporteurs Maina Kiai and Christof Heyns wrote that "[a]utonomous weapons systems that require no meaningful human control should be prohibited".[22] I would add that it is not only the case that autonomous weapons systems might circumvent meaningful human control; the greater concern is that they could render it impossible. The judgment required for effective and legal 'action according to rule' requires time for the assessment of a current situation, and decreasing timeframes due to increasing automation close down the time available for assessment. The proposed solution of 'human-machine teaming', moreover, is only effective to the extent that system designs maintain the system dynamics (more colloquially, the time) required to allow for meaningful human control.[23] This requirement, in turn, poses further limits to weapons system automation.

20 Assertions that "technology may evolve and meet the requirements [for human target identification] in the future" (cited in the Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), CCW/MSP/2/015/3, 2 June 2015, p. 14), or "[a]utonomous technologies *could* lead to more discriminating weapons systems" (ibid., p. 15, emphasis added) do not comprise evidence-based statements of fact.

21 Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns, and Special Rapporteur on the Rights to Freedom of Peaceful Assembly and of Association, Maina Kiai, for the Office of the High Commissioner for Human Rights, A/HRC/31/66, 4 February 2016, https://t.co/hpkjz7CfyV.

22 Ibid., para. 67.

23 See P. Scharre, Autonomous Weapons and Operational Risk, Center for A New American Security, 2016, https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk.

Implications for lawful weapons autonomy

Conceptual clarity regarding the capacities that enable situational awareness in the case of human combatants, with a particular focus on the principle of distinction, clarifies in turn the requirements for autonomous technologies, and more specifically for LAWS. Citing Article 48 of the First Additional Protocol to the Geneva Conventions, Crootof observes that one implication of the principle of distinction is that:

> *"parties are prohibited from using inherently indiscriminate weapons, which are usually defined either as weapons that cannot be directed at lawful targets or as weapons whose effects cannot be controlled. Additionally, any given attack in an armed conflict cannot be indiscriminate: it must be directed at a lawful target and cannot utilize indiscriminate weapons or methods of warfare."*[24]

The defining question for LAWS is whether the discriminatory capacities that are the precondition for legal killing can be reliably and unambiguously encoded in weapons systems. As noted above, this judgment has become increasingly difficult for human combatants, for several reasons. First, the conditions of so-called irregular warfare have removed traditional designations of battle zones and combatants, requiring much more subtle and uncertain readings of the presence of an imminent threat.[25] Second, because military systems involve increasingly complex, distributed, real-time networks of information and communication, the possibilities have amplified not only for strategic accuracy, but also for noise.[26] And finally, the intensifying automation of warfare has effected a progressive narrowing of timeframes for situational assessment.

24 A. Crootof, The Killer Robots Are Here: Legal and Policy Implications, 36 Cardozo Law Review 1837 (2015) 1873).

25 M.L. Flagg, Deputy Assistant Secretary of Defense at the U.S. Office of the Undersecretary of Defense for Acquisition, Technology and Logistics' research directorate, imagines a situation in which "a robotic system is in a battle zone, knows the mission, has been thoroughly tested, has the kinetic option, and its communications links have been cut off", and then asks whether that machine should then make the decision to deploy a weapon independently. But it is precisely this clarity that is absent in actual situations of war fighting; see S. Magnuson, Autonomous, Lethal Robot Concepts Must Be 'On the Table', DoD Official Says, National Defense Magazine, 3 March 2016, http://www.nationaldefensemagazine.org/blog/Lists/Posts/Post.aspx?ID=2110.

26 On the intransigence of this problem see for example P. Cronin, The Impenetrable Fog of War: Reflections on Modern Warfare and Strategic Surprise, Westport, CT 2008.

Lawand proposes that "[a] truly autonomous weapon system would be capable of searching for, identifying and applying lethal force to a target, including a human target (enemy combatants), without any human intervention or control".[27] But in the parenthetical 'enemy combatants' lies the crux of the problem: how is the identification of 'human target' with 'enemy combatant' confirmed? And what uncertainties characterise the category of 'enemy combatant' that confound, rather than clarify, the problem of legitimate target identification in contemporary warfare? Autonomous systems can be made reliable only to the extent that their environments, the conditions of their operation, can be specified, engineered and stabilised; these requirements do not hold in situations of combat.[28] All of the evidence to date indicates that this is at best an unsolved problem for machine autonomy, and at worst (and this is my position, for the reasons set out above) an unsolvable one.

In sum:

1. We take as our definition of lethal autonomous weapons such robotic weapons systems that, once activated, can select and engage *human* targets without further intervention by a human operator.
2. Autonomy in human or machine systems implies self-directed action, including, crucially in the case of military operations, 'action according to rules'.
3. 'Action according to rules' in the case of human action presupposes capabilities that the rules cannot fully specify; in particular, those competencies that are required to map the conditions assumed by the rule to actually occurring situations.
4. In the case of LAWS that would be adherent to IHL, machine autonomy requires reliable, unambiguous translation of rules for situational awareness, particularly for the identification of legitimate human targets (the principle of distinction), into machine-executable code.

27 K. Lawand, Fully Autonomous Weapon Systems, Statement, International Committee of the Red Cross, 25 November 2013, http://www.icrc.org/eng/resources/documents/statement/2013/09-03-autonomous-weapons.htm.

28 It is widely recognised that "as the behavior of automated systems becomes more complex, and more dependent on inputs from environmental sensors and external data sources, the less predictable they become", Asaro (n 15). And as the latter expert's panel observed, "[d]eploying a weapon system with unpredictable effects creates a significant risk of a breach of International Humanitarian Law", Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), CCW/MSP/2/015/3, 2 June 2015, p. 15.

5. Contrary to assertions regarding the rapid advance of artificial intelligence and robotics, there is no empirical evidence of progress in operationalising the capacities of situational awareness that are required for adherence to the principle of distinction.
6. This raises serious doubts regarding the feasibility of lethal autonomous weapons adherent to IHL.

To conclude, conceptual clarity regarding the capacities that enable situational awareness in the case of human combatants, with a particular focus on the principle of distinction, clarifies in turn the requirements for lethal autonomous weapons systems. The defining question for autonomous weapons is whether the discriminatory capacities that are the precondition for legal killing can be reliably and unambiguously encoded. My argument is that they cannot, and that as a consequence lethal autonomous weapons are in violation of IHL, and should be prohibited.

A Framework of Analysis for Assessing Compliance of LAWS with IHL Precautionary Measures

Kimberley N. Trapp*

Introduction

The Additional Protocols to the Geneva Conventions[1] were negotiated at a time of relative technological simplicity. Compliance with the prohibition against indiscriminate attacks,[2] supported by the obligation to take precautionary measures in planning and deciding to launch an attack,[3] was measured in terms of human effort – in gathering and assessing information about targets and their circumstances, and in taking critical decisions based thereon. In the current 'Information Age', some of that human effort has been replaced by machines – in that relevant data can be gathered (by surveillance technology), assessed (in part), and disseminated remotely by computers at a rate and volume that would have been science fiction in 1977.

* Senior Lecturer in Public International Law, UCL, Faculty of Laws. The framework of analysis set out in the first two sections of this paper draws on K.N. Trapp, Great Resources Mean Great Responsibility: A Framework of Analysis for Assessing Compliance with API Obligations in the Information Age, in Dan Saxon (ed.), International Humanitarian Law and the Changing Technology of War, The Hague 2013, p. 153.

1 This Chapter will focus on obligations as framed in the Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (API), 8 June 1977, 1125 UNTS 3. The ICRC and states which are not party to API, however, consider the obligations that are the subject of this study to reflect customary international law. See e.g. J.-M. Henckaerts and L. Doswald-Beck, Customary International Humanitarian Law, Vol. I: Rules, Cambridge 2005, Rules 15 and 16; M.J. Matheson, Remarks, 2 American University Journal of International Law and Policy 419 (1987) 423-426.

2 Indiscriminate attacks are prohibited under API. They are defined in part as attacks "which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated", Article 51(5)(b) API. This conception of 'indiscriminate' is referred to in terms of the 'proportionality' of an attack, and will be referred to as such throughout this paper.

3 In particular, Art. 57(2)(a)(i) API requires states to "do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects and are not subject to special protection but are military objectives [...] and that it is not prohibited by the provisions of this Protocol [including obligations regarding the proportionality of attacks] to attack them".

But technologically advanced states potentially aspire to move past this Information Age to an 'Age of Autonomous Weapons' – further cutting human effort out of the IHL compliance calculus. These technological developments – unforeseen at the time the relevant treaty standards were negotiated – raise difficult legal questions: how might we assess the IHL compliance of lethal autonomous weapons systems (LAWS)? Are the standards of IHL compliance sufficiently flexible to respond to the rate of technological development in the modern era, particularly where such development puts humans 'out of the loop' in reference to critical decision making functions? The answer to these questions is, as one might expect, complicated, and requires somewhat of an 'onion peel' approach.

There are in effect three layers of assessment: the outermost layer is the general international law standard applicable to the obligation to take precautionary measures under Article 57 Additional Protocol I (API). As explored in the first section below, this obligation is an obligation of conduct, not result – which is to say that compliance is measured in terms of *diligent* efforts made, not outcomes. The middle layer (explored in the second section) is informed by the general international law standard and involves a more specific assessment of compliance standards which address technological development generally, and obligations which involve the gathering and assessment of information in particular. Finally, the core of the relevant analysis involves specific consideration of the implications of LAWS in terms of IHL compliance.

Precautionary measures as an obligation of due diligence

Article 57(2)(a)(i) API requires states to do everything feasible to verify that a target is a military objective and that it is not otherwise prohibited by API to launch the attack (in particular, that the attack would not be disproportionate), before proceeding. While 'everything feasible' sounds like a rather high standard, the obligation is nevertheless understood as an obligation of conduct, not one of result.[4] Indeed, the Commentary to API notes that 'everything feasible' is understood in terms of "everything that was practicable or practical-

4 The distinction between obligations of conduct and obligations of result is derived from the Civil Law tradition and turns on an analysis of whether the primary rule requires absolutely that state conduct produce a certain result (obligation of result), or whether it requires only that a state make certain efforts to produce a desired, but uncertain, result (obligation of conduct). See e.g. ICJ, Case Concerning the Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro), Judgment, 2007, ICJ Rep 43, para. 430, for an application this distinction.

ly possible",[5] making it clear that the obligation to take precautionary measures is understood in terms of the efforts made. An assessment of compliance with the obligation to take precautionary measures must therefore focus on the process of verification and collateral damage assessment,[6] rather than outcomes.

Obligations of conduct, unlike obligations of result, are subject to a due diligence standard – and diligence, as a matter of international law, involves an 'available means' analysis. As a result, international jurisprudence and doctrine have highlighted the importance of accounting for available resources in assessing compliance with obligations of conduct (particularly obligations to develop a capacity to keep being informed, as discussed below).[7] The implication of conditioning the obligation to take precautionary measures on feasibility, practicability, and diligence is that an assessment of compliance will turn (to a certain extent) on the technological means available to belligerents. While this was an accepted consequence of the way in which the obligations were framed,[8] it does mean that parties to an armed conflict which *could* do more (account taken of the state of their technological advancement and available resources) cannot get away with implementing the lowest common denominator of precautions simply because their adversaries are not in the same technologically privileged position. This result does not undermine the reciprocal nature of IHL obligations – which apply equally even though compliance is assessed relative to particular capacity as a matter of law – but may well be the cause of some resentment and dissatisfaction in asymmetrical conflicts.[9]

5 ICRC, Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949, www.icrc.org/ihl.nsf/WebList?ReadForm&id=470&t=com, para. 2198.

6 Committee Established to Review the NATO Bombing Campaign against the Federal Republic of Yugoslavia [ICTY Committee of Experts], Final Report to the Prosecutor, (2000) 39 ILM 1257, para. 29.

7 See K.N. Trapp, State Responsibility for International Terrorism, Cambridge 2011, §3.1, for further detail.

8 Commentary to API (n 5), para. 2199.

9 See M.N. Schmitt, War, Technology, and International Humanitarian Law, Occasional Paper Series, Program on Humanitarian Policy and Conflict Research, Harvard University, July 2005, http://www.hpcrresearch.org/sites/default/files/publications/OccasionalPaper4.pdf, pp. 2-3.

Elements of 'everything feasible' in the information age

'Due diligence' is the general international law standard against which assessment of compliance with the API obligation to take precautionary measures is to be measured. In the specific circumstances of the obligation to do 'everything feasible' to verify, due diligence takes on a very particular form and is informed by technological development. While the API obligation to take precautionary measures was articulated at a time when 'feasibility' and expectations of civilian casualties would have been heavily conditioned by available technology, the obligation to take precautionary measures is nevertheless framed in language flexible enough to account for exponential technological advancements.

At the time of writing, these advancements have principally been in reference to information gathering, assessment and dissemination capabilities. In the Information Age, at least militarily developed states have access to a vast amount of information about the particular circumstances of targeted military objectives, can appraise changed circumstances in real time as a result of their persistent surveillance capabilities (resulting from information gathered by unmanned aerial vehicles (UAVs) and satellite imagery), and are developing the networking capabilities to disseminate critical information quickly to relevant (human) actors. Compliance with the API obligation to do 'everything feasible' is therefore the product of a functional partnership between machine and human operators – with the ultimate exercise of judgment and discretion as to the proportionality of an attack, and indeed the decision to attack, being the responsibility of human actors. And due diligence – measuring the efforts made to verify that an objective is military and to account for 'expected' incidental losses in launching an attack – will turn on an assessment of two separate obligations.

First, there is an obligation to use available means to develop relevant information gathering, analysis, and dissemination capabilities. The commentary to API notes the "*duty* of Parties to the conflict to have the means available to respect the rules of the Protocol".[10]

The second obligation relevant to compliance with the API obligation to take precautionary measures is an obligation to put available technologies and gathered information to good and diligent use.[11] As discussed further below, both the obligation to develop relevant information technologies and the obligation to put those technologies to good and diligent

10 Commentary to API (n 5), para. 1871 (emphasis added).

11 Obligations which require an assessment of (and appropriately tailored action based on) information have long been considered to be composite obligations – consisting of the two equally important obligations to develop capacity (to gather/assess/disseminate information) and diligently utilise developed capacity and its fruits; see Trapp (n 7), para. 3.11 and 3.12.

use are obligations of conduct and subject to a diligence assessment. If it were otherwise, states with limited resources would be held to the technologically advanced standards of developed states in terms of their information gathering, analysis, and dissemination capabilities, and this would effectively guarantee a breach of the obligation to take precautionary measures – no matter how diligently such states put their more limited information gathering and dissemination capabilities to use.

Obligation to develop relevant capabilities

Specifically, measuring diligent compliance with the obligation to develop relevant capabilities will turn on a factual analysis of the socio-political circumstances of the relevant state, its general technological capabilities, and its particular development of technology relevant to battlefield identification and assessment of potential targets. For instance, states with limited resources, including relatively small intelligence budgets, or states which are not perpetually engaged in (or threatened with) armed conflicts, are unlikely to have developed persistent surveillance capabilities. To the extent that such states find themselves in a situation of armed conflict, they would certainly need to act diligently in adapting their limited existing capabilities to battlefield purposes in order to meet their API obligations. On the other hand, states which already have extensive persistent surveillance capabilities and are engaged in long-term armed conflicts would not be acting diligently if some development of those capabilities were not aimed specifically at, for instance, effective target identification.

The obligation to diligently develop relevant information capacity itself has three elements, as explained in the following.

c. Gathering information

States with relevant technical resources, in particular states which are at war for an extended period of time, would be expected to develop 'just-in-time' capabilities which enable them to gather information concerning the nature and circumstances of a target whenever needed.[12] Such technological developments, coupled with more traditional reliance on human intelligence and reconnaissance, amount to a diligent effort to develop information gathering capabilities relevant to meeting the API obligation to take precautionary measures.

12 See e.g. R. Best, Intelligence, Surveillance, and Reconnaissance (ISR) Acquisition: Issues for Congress, Congressional Research Service, December 2011, p. 8.

d. Timely analysis

Any assessment of a state's efforts to develop a capacity to analyse gathered information needs to be realistic – accounting for the volume of raw data collected, and the analysing, interpreting, and integrating *all* that data in a timely manner, which is currently still practically impossible, given much of the analysis is still carried out by human operators. A realistic assessment of a state's efforts to develop relevant and time-effective analysis capabilities also needs to be highly sensitive to competing priorities. This is because international law is silent regarding the way in which a state allocates its resources, and indeed must be so given the ever-increasing extent of international regulation and concomitant demands on limited financial, technical, and human resources. Diligence therefore needs to be measured on the basis of a state's efforts to *improve* on timely analysis, appreciating the impossibility of entirely overcoming the limitations of human ingenuity and the current limitations of 'artificial intelligence'.

e. Dissemination and usability

While timely production of intelligence may always be a difficulty given the vast amount of raw data collected, the timely dissemination of such intelligence, once analysed, and its usability is a key area of military capability development. Again, precautionary measures require diligence in this regard – continued efforts to respond to the circumstances of armed conflict with a view to maximal IHL compliance.

f. Conclusion

Diligence requires that a state lives up to the challenges of the armed conflicts in which it is engaged (for instance, the challenges of asymmetrical warfare) – particularly as regards the necessity of properly identifying targets (whether immovable, movable or human), and that its information technology development strategy is responsive to those challenges. An important part of diligent development is a good faith effort to rectify identified inadequacies, which in turn depends on the respective state not turning a blind eye to technical difficulties encountered by its military in the field. This might be considered the 'lessons

learned' feature of diligent development[13] – requiring technological development which responds precisely, but subject to competing budgetary priorities, to the need for accurate, timely, and actionable intelligence.

Obligation to put developed capabilities to good and diligent use

For the purposes of assessing diligent compliance with the API obligation to take precautionary measures in the Information Age, it is assumed that critical decisions are taken by human operators. And such human operators, for example military commanders, are held to a standard of reasonableness in their critical decision-making, in particular decisions as to whether a target is indeed a military objective, and any proportionality calculus necessitated by the circumstances.[14] Any assessment of individual compliance with the obligation to take *feasible* precautionary measures, including in reference to a military commanders' *expectation* of civilian losses, will necessarily draw on the level of technological advancement of the state. This is because compliance is assessed on the basis of information *reasonably* available *in the circumstances*,[15] and such circumstances will to a large extent be driven by a state's development of information gathering, analysis, and dissemination capabilities – coupled with certain temporal factors in respect of the particular attack.[16]

13 The 'lessons learned' feature of an obligation to do 'everything feasible' is implicit in the NATO Bombing Report's conclusion that states can rely on a proven track record of distinguishing between military objectives, and civilians and civilian objects; see NATO Bombing Report (n 6), para. 29. *E contrario*, where information gathering, assessment, and dissemination methods have resulted in several mistakenly identified military objectives, diligence requires a state to re-evaluate its processes and address any inadequacies.

14 ICTY, Prosecutor v Galic, IT-98-29-T, Trial Chamber Judgment and Opinion, 5 December 2003: "it is necessary to examine whether a reasonably well-informed person *in the circumstances* of the actual perpetrator, making reasonable use of the information *available* to him or her, could have expected excessive civilian casualties to result from the attack" (emphasis added).

15 See the state practice reviewed in Henckaerts and Doswald-Beck (n 1), pp. 363-365.

16 See Trapp (n 7) for an analysis of temporal factors affecting an assessment of compliance with precautionary measures.

LAWS and API Compliance

The final layer of IHL compliance examined in this chapter concerns the way in which presently applicable standards of compliance (explored in the second and third sections above) adapt to further technological development – whereby information is assessed and actioned within a single weapons system, without human initiation or further intervention. This chapter will address the case of LAWS which operate at the highest levels of autonomy and with humans 'out of the loop', meaning that the weapons system performs critical functions such as target acquisition, tracking, selection, and attack without human initiation or intervention. The measure of interaction between human and machine, which is a feature of compliance with IHL in the Information Age, would thereby be limited to the time prior to programming and parameter-setting.

The framework of analysis set forth in the second section above assumes a functional partnership between machine and human operators – with the ultimate exercise of judgment and discretion as to the proportionality of an attack, and indeed the decision to launch an attack, being the responsibility of human actors. This feature of Information Age compliance with API obligations is precisely what is missing from LAWS, where analysis of available information, proportionality judgment based thereon, and kill decisions are folded into a single weapons system without a 'human in the loop'. The implications of this 'consolidation' of relevant tasks in assessing API compliance are set forth below.

Elements of API compliance fold into each other

In cases of a functional partnership between machines and human operators during an armed conflict, there is a logic to splitting API compliance into two separate elements, one in reference to capacity and technological development (which will facilitate compliance with API obligations), the other in reference to the use to which that capacity and technology is put by human operators (in compliance with API obligations). In the case of LAWS, however, the capacity/technology (and its development) is not separate from its end use – even if not all the information gathering is done within the one weapons system, the analysis of the gathered data and critical decision-making is at least part of the same weapons system network. The two separate elements of a diligence assessment in respect to compliance with API obligations in the Information Age therefore collapse one into the other in an Age of Autonomous Weapons. Assessing compliance with an obligation to do 'everything feasible' becomes a significantly more focused judgment – as everything depends on the development and testing of the technology.

This has clear implications for Article 36 API – which imposes an obligation on State Parties to "determine whether [a weapon's] employment would, in some or all circumstances, be prohibited" by the Protocol. In the Information Age, these assessments tend to focus on the precision of the weapon. In an Age of Autonomous Weapons, Article 36 assessments need to focus on the entire range of API obligations – and indeed it may be impossible to determine, *a priori*, whether LAWS can comply with the obligation to take precautionary measures. This is not least because artificial intelligence now in development would potentially allow for weapons systems to 'learn' from experience. Any weapons system subject to the rigours of Article 36 in laboratory conditions will therefore be different from, or rather an earlier version of, the weapons system later deployed in combat situations.

Decrease in margin of appreciation

The second implication of LAWS in assessing compliance with the API obligation to take precautionary measures is in reference to the margin of appreciation states enjoy in the Information Age. In particular, states enjoy a margin of appreciation in meeting the capacity development obligations inherent in due diligence obligations. As discussed above, due diligence obligations are conditioned by an 'available means' analysis, and the margin of appreciation is an important aspect of any such resource-based analysis. This is in part because states are faced with competing priorities, and how they manage their resources will to a large extent depend on the nature and number of threats they face, and the acuteness of any such threats. International law therefore has very little to say about how states should prioritise resource allocation.

However, in respect to LAWS with a high degree of autonomy as regards critical functions, the margin of appreciation that even technologically advanced states enjoy has to decrease. This is because everything depends on the technology in an 'everything feasible' compliance calculus – and that, of course, increases the standard against which capability development is measured, as 'diligence' is not shared between technological development and human end use.

That the margin of appreciation decreases, perhaps even significantly, is made clear in the relevant literature. For instance, in the ICRC Report produced after the 2014 round of expert meetings, one commentator suggested that if it is not possible to 'guarantee' that the weapons system would comply with IHL in all circumstances, then it would be

unlawful.[17] Such statements perhaps go somewhat farther than legally warranted (in that capacity development, as part of the precautionary measure obligation, is subject to a due diligence standard, while *guaranteeing compliance* is what would be required of an obligation of result). The standard of compliance needs to be adapted to technological developments between 1977 and the Information Age, and 'everything feasible' will have to respond to an increasingly *exclusive* technological environment. And it is clear that when everything depends on the technology, 'diligence' will require significantly more of states than is presently required.

Human in/out of the loop

The final implication of LAWS in assessing API compliance is in reference to the 'human in the loop' issue, again assuming a high degree of autonomy. For the foreseeable future, given the stage of artificial intelligence (AI) and artificial learning intelligence development, 'everything feasible' should be measured against a 'human in the loop' standard – precisely because it is always feasible for them to be kept 'in the loop', i.e. in control. Again, this increases the standard against which technology is measured when determining compliance with precautionary measures. Until machines can exercise judgment, and engage in a balancing act that even those who regularly do so might not be able to explain (much less program AI to do), the 'human in the loop' standard against which autonomous weapons compliance will be judged is simply too high for such weapons systems to pass an Article 36 API review.

Conclusion

While the obligation to take precautionary measures and to do 'everything feasible' is of universal application, assessment of compliance must have regard for the very particular circumstances of each party to an armed conflict. It is therefore true that great resources mean great responsibility. And states which have the resources available to even contemplate the development of LAWS have great responsibilities indeed.

17 ICRC, Expert Meeting, Report on Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects, Geneva, Switzerland, 26-28 March 2014, https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf, p. 22.

The flip side of great responsibility is that great resources engender great opportunities. In particular, the capacities developed in the Information Age (including to gather and disseminate information remotely) give states the freedom to counter the dangers to which their national forces are exposed in meeting API obligations to take precautionary measures. It is only right that states which have the technological means available to put their armed forces out of harm's way in conducting war should also be held to the standard of *feasibility* suggested by those technological means in protecting civilian populations from its ravages. In this regard, LAWS go one significant step further than current technologies – constituting the ultimate force protection weapon. And IHL simply does not allow for a state to prioritise protecting its own forces above protecting civilians. Given the increased standards of IHL compliance in reference LAWS as discussed above, it is difficult, or rather near impossible, to see a path to IHL precautionary measure compliance for LAWS.

Predictability and Lethal Autonomous Weapons Systems (LAWS)

Wendell Wallach *

Introduction

Does predictability provide an overriding concept and perhaps a metric for evaluating when lethal autonomous weapons systems are acceptable or when they might be unacceptable under existing international humanitarian law? Arguably, if the behaviour of an AWS is predictable, deploying it might be considered no different from, for example, launching a ballistic missile. This, of course, presumes that we can know how predictable the behaviour of a specific AWS will be.

LAWS will take many forms. They will range from dumb systems (e.g. landmines) to systems with limited intelligence, swarming drones, and eventually extremely intelligent systems. Blind autonomy (as in landmines) has already been rejected by civilised nations through the Ottawa Treaty. All systems can fail in unpredictable ways, but commonly this causes an inability to function. On occasion, however, failure results in unpredicted behaviour. The unpredictability of behaviour increases with the intelligence, mobility, and learning capability of a system. In the deployment of an AWS, the question that must be addressed is whether or when unpredictable behaviour poses an unacceptable level of risk. What level of risk would be unacceptable? Can military analysts actually know with any degree of certainty the level of risk an AWS poses in a specific context?

The debate as to whether LAWS will or will not violate IHL has to date been piecemeal, circular, and often repetitive as proponents and opponents of a prohibition put forward arguments and counterarguments. The search for definitions of key terms such as autonomy, meaningful human control, and predictability suggest perhaps a maturation of that discussion. However, we should not be surprised when any definition fails to be fool-proof, meaning it fails to establish a bright line as to which weapons system or use of a weapons system is acceptable or unacceptable. There will always be ambiguous situations on the margins for acts of aggression that violate international norms. A central challenge confronting CCW is whether those ambiguities will or should be used as justification for a failure to act decisively.

* Yale Interdisciplinary Center for Bioethics.

Calls to prohibit AWS fall into two broad arguments:

1. Ethical and legal positions that machines should not make life and death decisions, often based upon competence arguments that intelligent systems fail to have the requisite capabilities for making discriminating and proportional decisions as required under IHL.
2. Risk concerns that the behaviour of autonomous weapons will be unpredictable and therefore subject to behaviour that will violate IHL, even when this is not the intention of those who deploy the system.

This paper will address the second of these arguments – whether offensive LAWS that can select a target and kill people are low risk and will exhibit adequate predictability in their behaviour.

Autonomy, safety, control, reliability, predictability, and testing

Ensuring that autonomous computer systems and robots are truly beneficial and safe is a deeply problematic challenge, even for domestic applications and non-combat related military applications. Because we witness remarkable feats performed by computer applications, it is easy to lose sight of the time, expense, and difficulties entailed in fixing errors in their programming code, and the countless billions of dollars spent by countries and companies on applications that were abandoned. Furthermore, most applications do not cause any physical harm to people. In March 2015, when AlphaGo beat the world's best Go player in a match,[1] only the pride of the player, and perhaps the vanity of our species' belief in its superiority suffered.

The failure of mission-critical software applications has killed people, destroyed property, and bankrupted industries and financial institutions. Commonly, such failures were tolerated because they were either unexpected or the benefits far outweighed the costs. Even new applications, such as self-driving cars, which are anticipated to reduce traffic fatalities significantly, will in some situations kill people that human drivers might not have killed. Driving is a social activity that involves solving problems and making decisions that cannot all be programmed into a computer system in advance.

1 C. Moyer, How Google's AlphaGo Beat a Go World Champion, Wired, 28 March 2016, http://www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611/.

While successful computational systems automate both routine and complicated activities, autonomous systems will be increasingly more adaptive, select among various courses of actions, and have some learning capabilities. Autonomous systems are best understood as complex adaptive systems that occasionally act unpredictably, have tipping points that lead to fundamental reorganisation, and can even display emergent properties that are difficult, if not impossible, to explain. Learning algorithms and other forms of artificial intelligence (AI) are becoming commonplace as integral components of a wide array of applications. Over the past year, AI engineers have expressed deep concern about their limited ability to ensure the safety and control of increasingly autonomous learning systems, or effectively test and verify the systems' purported benefits. Thus more than 8,600 AI researchers (as of March 2016) have signed a Future of Life Institute open letter and accompanying document entitled, 'Research Priorities for Robust and Beneficial Artificial Intelligence', which calls for research to begin on developing strategies, algorithms, and other tools to ensure that AI systems will be beneficial, robust, safe, and controllable.[2] While technological optimists assume that all technological problems can be solved, many AI researchers privately admit that if the safety of increasingly autonomous systems remains in doubt, the pursuit of many applications of AI should be relinquished. While some of their fears relate to the speculative futuristic possibility of creating a form of superintelligence that is unfriendly to humans, the problems posed by poorly tested and unpredictable autonomous systems are already evident for the semi-intelligent systems available today.

None of the above concerns are explicitly about lethal autonomous weapons systems. Introducing autonomy into weapons systems adds the additional worry that unpredictable behaviour might start a war, exacerbate hostilities, or be a violation of IHL. Pondering the short-term benefits of AWS, military strategists may conclude that the risks are warranted. However, CCW must decide whether any increase in IHL violations should be accepted, even when the same class of weapons systems might, for example, decrease civilian casualties in a different context. While the concept of *proportionality* suggests a consequentialist calculation, generally IHL is not viewed as merely a trade-off between risks and benefits. At the very least, we (i.e. policy makers and theorists) must find a way to strengthen *jus in bello* for all contexts. The question before the CCW should be whether we can "establish a set of incentives that drive the creation of acceptable technologies given the environment within which we propose to use them".[3]

2 Available online at http://futureoflife.org/ai-open-letter/.
3 Modified from an email sent to the author by Braden Allenby.

Predictability

Nothing less than a law of physics is absolutely predictable, and even gravity is a contingent law, waiting to be falsified by the observation of an apple that floats up off a tree.[4] Thus there are only degrees of predictability.

In classical theory, known causes lead to analytical predictability. However, statistics and quantum physics rely upon probabilities to represent the distribution of possible events and the likelihood of any particular occurrence. Mirroring this distinction, the military strategist Antonie-Henri Jomini emphasised analytical predictability through adding up elements such as the number of guns and soldiers, while Carl von Clausewitz stressed the importance of variable elements, such as the emotions or psychological state of combatants, and their interactions in undermining simple predictability.[5]

The second attribute of military action is that it must expect positive reactions, and the process of interaction that results. Here we are not concerned with the problem of calculating such reactions – that is really part of the already mentioned problem of calculating psychological forces – but rather with the fact that the very nature of interaction is bound to make it unpredictable.[6]

According to Alan Beyerchen:

> *"[I]n a profoundly unconfused way, [von Clausewitz] understands that seeking exact analytical solutions does not fit the nonlinear reality of the problems posed by war, and hence that our ability to predict the course and outcome of any given conflict is severely limited."*[7]

However, in unifying the dynamic theory of war posed by von Clausewitz with the more analytical Jomini, it is important to note that both presumed that the tools of war, the weaponry itself, is reliable and not a significant source of unpredictability.

4 Another example: any proof that dark energy and dark matter do not exist would be a falsification that gravity is a law applicable in all situations.

5 C. von Clausewitz, On War, ed. and trans. by M. Howard and P. Paret, Princeton 1976.

6 Ibid., p. 139.

7 A.D. Beyerchen, Clausewitz, Nonlinearity, and the Unpredictability of War, International Security, Winter 1992/1993, http://www.clausewitz.com/item/Beyerchen-ClausewitzNonlinearityAndTheUnpredictabilityOfWar.htm, pp. 59-90.

The degree of predictability or unpredictability can in theory be represented as a probability. The probability that predictions based upon the law of gravity will be accurate approaches 100 percent, lacking known proven exceptions. Nevertheless, even the law of gravity can lead to surprising behaviour, such as weightlessness.

With the flipping of a coin we cannot predict an individual event or succession of events in advance, but at least we statistically understand the probabilities. The probability that a coin toss will yield a head is 50 percent for each incident. Given that, should one bet against a succession of ten heads in a row? Most people dramatically underestimate the likelihood of ten heads in a row. On average ten successive heads will occur once in every 1024 flips. Now consider a computer that flips a simulated coin once every millisecond (thousandth of a second). It will produce a series of ten heads roughly once every seconds. The computer simulation does not change the probability of the individual flip (50 percent), but it does speed up the rate (contract the time) over which a low probability event will occur.

In the conduct of warfare, the predictability of weaponry means that within the task limits for which the system is designed, the anticipated behaviour will be realised, yielding the intended result. Nevertheless, an unanticipated event, force, or resistance can alter the behaviour of even highly predictable systems. In the non-linear dynamics of contemporary chaos theory, under certain circumstances even a subtle influence can have a dramatic effect (commonly known as the 'butterfly effect').[8]

There is a fundamental difference between a truly unanticipated event or influence upon a highly predictable system's behaviour, and a low probability event that can have a high impact and far-reaching consequences. Investors, bankers, and military strategists who fail to distinguish between the two are destined for failure. An investment firm that follows the same high probability strategy year in and year out may be profitable for decades or even centuries, and then in one year go bankrupt with the advent of a low-probability/high impact event, what the theorist Nassim Taleb labelled a 'black swan'.[9]

Good military planning raises the likelihood (prediction) of a successful campaign. Planning helps control what can be controlled through training, discipline, and testing. Success also requires tactical and strategic planning that covers as wide an array of recognised

8 The concept of the 'butterfly effect' was first introduced by Edward Norton Lorenz in a 1972 lecture entitled 'Predictability: Does the Flap of a Butterfly's Wings in Brazil Set Off a Tornado in Texas?', online at http://eaps4.mit.edu/research/Lorenz/Butterfly_1972.pdf.

9 N.N. Taleb, The Black Swan: The Impact of the Highly Improbable, New York 2007.

contingencies as possible, and adaptability or flexibility when making real-time choices. Robust command and control is essential, as is the reliability of the increasingly sophisticated equipment upon which the modern military depends.

From reliability to predictability

Reliability, and similar engineering terms, has traditionally served as a standard for evaluating mechanical systems and weaponry. Testing can provide a degree of certainty about the reliability of purely mechanical or automated systems, including systems that make simple straightforward computational calculations. The reliability of older, poorly maintained equipment will go down, even while the exact moment of failure can often not be predicted in advance. Through the 1950s, an automobile tire blowout was a common occurrence. However, what happened after a blowout was unpredictable, but drivers could be trained to handle most blowouts successfully. Better design and materials combined together with good maintenance strategies mean that today blowouts seldom occur.

A rifle, or similarly engineered system, is certified as reliable within specified parameters after extensive testing. But once rifles are deployed in the field, an experiential process takes over to refine the understanding of the weapon and its successful use. For example, during the Vietnam War it became well known among soldiers and officers that the M-16 was more prone to jamming and problems under field conditions than the AK-47.

The need for a definition of 'predictability' has only arisen because traditional engineering standards of reliability no longer suffice when evaluating the behaviour of complex dynamic systems, increasingly autonomous computers and robots, and computational systems with limited but expanding artificial intelligence and leaning capabilities. Fully automated systems are predictable in the sense that managers know what the system will do. An AI system is unpredictable because one cannot always know what it will do. For our discussion, the need for a definition of predictability rests upon the recognition that the increasingly complex hardware and software that will support many forms of AWS will on occasion produce behaviour that could not be predicted in advance and may be surprising, if not dangerous, in its effect or impact. Any definition of predictability will be fungible and entangled with definitions for other terms including reliability, certainty, trustworthiness, meaningful human control, testability, verifiability, and traceability.

Establishing a useful metric for determining predictability will not be easy, but one possible source is the use of the concept of 'uncertainty' in cost-benefit analysis (CBA) and in risk assessment (RA). The methodologies used for assessing costs, benefits, and risks of technology

projects that can have a societal impact or alter environments has required representations of 'uncertainty' to build adequate quantitative tools useful for evaluating the acceptability of the proposed projects. While CBA and RA provide rich methodologies, there is controversy as to the values inherent in these tools and whether either is sufficient for determining public policy as to the acceptability of any project analysed.

Software reliability

In software engineering, the predictability of the product (i.e. the software) refers to methods for optimising and quantifying aspects of software quality. Two important questions regarding the predictability of a piece of software revolve around the issue of correctness (how often does the software execute according to its specifications?) and reliability (roughly, how often does it perform correctly in a particular environment without crashing?). Software safety routines can be introduced into the system design to ensure that it cannot *melt down*, but these do not by any means guarantee reliability.

The goal is predictability, but that goal is elusive. Systems dependent on sophisticated software are intrinsically less reliable than purely mechanical artefacts. Billions of dollars have been spent on software systems that have been abandoned because they could not be relied upon. Debugging a state-of-the-art automated luggage sorting systems delayed the opening of Denver, Colorado's new International Airport for 16 months at a cost of $1.1 million a day.[10] As of 2013, an abandoned NHS patient record system cost British taxpayers around £12 billion.[11] The U.S. Air Force abandoned an expeditionary combat support system in 2012 after an expenditure of $1.03 billion.[12]

Unreliable software is commonplace, and even successful consumer products such as operating systems can be released with tens of thousands of known and unknown bugs. Bugs can lead to software and system failure, but they can also produce miscalculations. In 1994,

10 Calleam Consulting Ltd., Case Study – Denver International Airport Baggage Handling System – An Illustration of Ineffectual Decision Making, 2008, http://www5.in.tum.de/~huckle/DIABaggage.pdf.

11 D. Martin, £12bn NHS Computer System is Scrapped, Daily Mail, 22 September 2011, http://www.dailymail.co.uk/news/article-2040259/NHS-IT-project-failure-Labours-12bn-scheme-scrapped.html.

12 C. Kanaracus, Air Force Scraps Massive ERP Project After Racking up $1B in Costs, Computerworld, 14 November 2012, http://www.computerworld.com/article/2493041/it-careers/air-force-scraps-massive-erp-project-after-racking-up--1b-in-costs.html.

a calculation flaw in an entire line of Pentium 486 chips had to be replaced by the Intel Corporation at a pre-tax charge of $475 million.[13] The online trading firm Knight Capital Corp. lost $440 million in 45 minutes due to the introduction of faulty software in August 2012.[14]

Debugging software is a time-consuming and costly process, and often only done in a satisfactory manner by successful companies and governments with adequate budgetary resources. Even good companies can release poor products (e.g. the initial introduction of Apple Maps in 2012).

Software failures are truly commonplace, and should not be viewed as an anomaly. They can even lead to deaths. On 25 February 1991 during the Gulf War, a software delay led to the failure to track and intercept an incoming Scud missile in real time. According to the U.S. Government Accounting Office, 28 were killed and more than 100 were injured.[15]

Software reliability is also dependent upon the quality of the data available to it – "garbage in, garbage out." On 7 May 1999, the U.S. mistakenly bombed the Chinese embassy in Belgrade during a NATO operation. According to testimony, the wrong coordinates had been selected for the strike, which was aimed at a Yugoslav Federal Directory for Supply and Procurement.[16] In this case, the error and oversight was that of humans, but the bad coordinates could have just as easily been fed into the computer of an AWS.

After decades of experience in software engineering, David Parnas concluded that software is intrinsically unreliable, that the techniques commonly used to build military software fail to be robust enough to assemble complex systems, that research in artificial intelligence is unlikely to be helpful in building reliable military software, and "that military funding of research in software and other aspects of computing science is inefficient and ineffec-

13 L. Flynn, The Executive Computer; Intel Looks Beyond the Pentium, The New York Times, 26, February 1995, http://www.nytimes.com/1995/02/26/business/the-executive-computer-intel-looks-beyond-the-pentium.html.

14 J. Strasburg and J. Bunge, Loss Swamps Trading Firm, The Wall Street Journal, 2 August 2012, http://www.wsj.com/articles/SB10000872396390443866404577564772083961412.

15 N.L. Lewis et al., Casualties and Damage from Scud Attacks in the 1991 Gulf War, MIT Center for International Studies, 1993, http://web.mit.edu/ssp/publications/working_papers/wp93-2.pdf.

16 S.L. Myers, Chinese Embassy Bombing: A Wide Net of Blame, The New York Times, 16 April 2000, http://www.nytimes.com/2000/04/17/world/chinese-embassy-bombing-a-wide-net-of-blame.html.

tive".[17] Parnas was writing to explain his decision to resign from the Panel on Computing in Support of Battle Management in 1985, convened by the U.S. Strategic Defense Initiative Organization, but his insights are as applicable today as they were thirty years ago.

Security and Cyberwarfare

Poor security, hacking, and attempts by bad actors to game or outsmart the software can of course undermine the reliability of computer systems. It is foolhardy to underestimate the vulnerability of computerised weaponry in an age of cyberwarfare, when systems can be hacked and even redirected.

Complex adaptive and learning systems

Systems theory provides a starting point for studying highly networked technological artefacts that are deployed in rapidly changing and inherently unpredictable environments. As mentioned above, autonomous systems are best understood as complex adaptive systems. Complex adaptive systems cannot only act in unanticipated ways when confronted with new situations and new inputs, but they also have tipping points, and exhibit emergent properties that are difficult to explain. Complex systems fail for a variety of reasons including (but not limited to): incompetence or wrongdoing, design flaws or vulnerabilities, underestimating risks and failure to plan for low probability events, and 'black swans' – unforeseen low probability events. There are also what Charles Perrow referred to as 'normal accidents', when no one did anything wrong and yet the system fails.[18] A classic example of a 'normal accident' was the near meltdown of a nuclear reactor at Three Mile Island in Pennsylvania on 28 March 1979. Three components failed at the same time. While there were backups for the failure of each component, no one had considered a simultaneous failure of those three components. Furthermore, it would have been impossible to consider all the combinatorial failures of different components in a complicated and complex system such as a nuclear reactor. After reviewing the literature analysing the failure of Three

17 D.L. Parnas, Software Aspects of Strategic Defense Systems, 28 Communications of the ACM 1326 (1985), https://pdfs.semanticscholar.org/26ae/89f725ede99590b3a63e0780872c32a6871a.pdf; see also D.L. Parnas et al., Evaluation of Safety-Critical Software, 33 Communications of the ACM 636 (1990), http://www.cs.unm.edu/~cris/591/parnas1990evaluation.pdf.

18 C. Perrow, Normal Accidents: Living with High-Risk Technologies New York 1984. The list of reasons why complex adaptive systems fail is more fully developed in my book: W. Wallach, A Dangerous Master: How to Keep Technology from Slipping Beyond Our Control, New York 2015.

Mile Island and other disasters, the popular writer Malcolm Gladwell concluded that, "we have constructed a world in which the potential for high-tech catastrophe is embedded in the fabric of day-to-day life".[19]

Since Perrow first proposed his theory of 'normal accidents', it has been fleshed out into a robust framework for understanding the safety of hazardous technologies. 'Normal accident theory' has been contrasted to 'high reliability theory', which offers a more optimistic model for strategic planning.[20] Arguably, good strategic planners would evaluate their proposed campaigns under the assumptions of both 'high reliability theory' and 'normal accident theory'. However, such comparisons will produce dramatically contrasting visions of the likelihood of success.

Learning algorithms have become commonplace as software features of increasingly sophisticated computational systems. With recent breakthroughs using deep-learning algorithms, learning systems are expected to be ubiquitous. Deep-learning is an approach that uses multi-layered neural networks exposed to massive amounts of data. Neural networks learn by creating their own system of classification, recognise correlations between elements, and strengthen or weaken connections between these neural nodes. However, a neural network cannot explain what it does, nor can its process be fully analysed by engineers.

Until recently, even computational learning approaches failed at basic problems such as perception, but deep-learning algorithms have, for example, been successful at labelling correctly all the objects in a picture. Google's 'DeepMind' used a deep-learning algorithm to develop AlphaGo, which recently won a match (4-1) against Lee Sodol, a world champion Go player. Given the importance of autonomous systems recognising the features of the environments they move through, such learning capabilities will be essential if autonomous systems are to demonstrate even a modest degree of situational awareness.

Testing

The performance of a complex adaptive system can be satisfactory and even predictable when used in a routine way within fairly uniform and stable contexts or environments. Nevertheless, the behaviour of both complex and learning systems is difficult to predict in

19 M. Gladwell, Blowup, The New Yorker, 22 January 1996, http://www.newyorker.com/magazine/1996/01/22/blowup-2.

20 See S.D. Sagan, The Limits of Safety: Organizations, Accidents, and Nuclear Weapons, Princeton 2013, p. 13.

advance when confronted with either new contexts or totally new inputs. Without testing, engineers can often not know how the system will behave. Even so, it can be impossible or too costly to test an autonomous system within every context, or every environment, it will encounter, or with all the combinations of possible inputs. Indeed, even a simple calculator has an exponentially untestable combination of inputs. Furthermore, each programming error fixed or new feature added can alter the behaviour of a complex adaptive system. No company or government can repeat all tests for each upgrade. Major software and hardware vendors commonly release upgrades once basic testing demonstrates satisfactory performance, and then wait for feedback from user experience to begin addressing the serious problems overlooked. Obviously, such trial by error is fraught with ethical and legal issues when addressing the riskiness of systems that can cause harm or kill people.

Learning systems are even more problematic. Each new task or strategy learned can alter a system's behaviour and performance in a variety of different contexts. In some cases, even additional information will alter behaviour. Retesting dynamic systems that are constantly learning is impossible.

Furthermore, learning is not just a process of altering information, it can alter the very algorithms that process the information. Placing a system on the battlefield that can change its programming raises the risk of unanticipated behaviour significantly.

Coordination Failures

It is easy to retrospectively blame an officer or operator for deploying a device in a situation where it might act in an unpredictable manner that violates IHL. However, it is unreasonable to expect that all officers will understand the performance limits of successive iterations of an LAWS, when even the engineers who built the system cannot fully understand the system's predictability in advance.

Over the coming decades, most of the autonomous systems deployed will continue to be joint cognitive systems,[21] that is, human operators will direct some activities and other activities will be performed by the systems with little or no human involvement. Human involvement, however, does not necessarily imply meaningful human control in the selection of targets and in the act of killing.

21 D. Woods and E. Hollnagel, Joint Cognitive Systems: Patterns in Cognitive Systems Engineering, Boca Raton 2006.

A semi-intelligent device will function as part of a team, performing some tasks independently, and other tasks as directed by human operators. This type of team exhibits a complex choreography between its human and non-human actors. Such complex systems depend upon a high level of coordination. Confusion can occur between the subroutine the autonomous system is attempting to follow, and what the human operators believe the autonomous system is attempting to do. This was the case on 6 December 1999 when after a successful landing, a Global Hawk unmanned air vehicle veered off the runoff and its nose collapsed in the adjacent desert, incurring $5.3 million in damages.[22]

When anything goes wrong, initially the operators are usually judged to be at fault, largely because of presumptions that the actions of the system are automated, while humans are presumed to be the adaptive players on the team. Commonly, the proposed solution to the failure of a joint cognitive system will be to build more autonomy into the computational system. Unfortunately, this does not solve the problem. Anticipating the actions of a smart system becomes more and more challenging for a human operator as the system and the environments in which it operates become more complex. Expecting operators to understand how a sophisticated computer 'thinks', and anticipate its actions so as to coordinate the activities of the team, actually increases the responsibility of the operators. Coordination challenges add one more way in which an increasingly autonomous system can fail, producing unpredictable and unanticipated behaviour.[23]

Unpredictability in AWS

Can we actually determine the degree of unpredictability of any AWS? What degree of unpredictability is acceptable in a weapons system?

Political leaders and military strategists may wish to believe that the LAWS they will deploy will have a high degree of predictability, but they have no means of ascertaining whether their confidence in an AWS is warranted. While predictability of the behaviour for some very simple AWS can be relatively high, at this stage in the development of autonomous systems there is no good method, short of truly extensive testing, for determining the degree of predictability in an autonomous system's performance.

22 M. Peck, Global Hawk Crashes: Who's to Blame?, 87 National Defense 594 (2003), http://www.nationaldefensemagazine.org/archive/2003/May/Pages/Global_Hawk3871.aspx.

23 D. Woods et al., Behind Human Error, 2nd ed., Surrey 2010.

Uncertainties

As a class of IT applications, autonomous systems come fraught with uncertainties: uncertainties as to the many different forms and applications that will be explored and deployed; uncertainties as to whether it can be known when they will fail or act in an unpredictable manner; uncertainties as to whether the degree of unpredictability in their behaviour can be ascertained; and uncertainties as to whether safety and controllability can be ensured.

Over time, increasingly sophisticated AWS will be deployed. Therefore, it behoves the member states of the CCW to not be short-sighted in their evaluation of what will be a very broad class of military applications. Indeed, as a weapons platform, lethal autonomy can be added to any and all munitions from a machine gun to an atomic bomb. Given even a very low degree of unpredictability, the CCW must not appear to greenlight autonomous systems that can detonate weapons of mass destruction.

Levels of risk

The risk that a bad event will occur is often quantified as the probability of the event multiplied by its consequences.[24] Clearly, a LAWS that functions as a platform for a machine gun has an immediate destructive impact that pales in comparison to an AWS that can launch a ballistic missile. Even when both systems have a low likelihood of performing unpredictably, the level of risk posed by the more powerful munitions is significantly higher.

Some inherently risky technologies are tolerated because the risk is truly low while the rewards can be high. Yet even in these cases, the likelihood of a tragic event can be lowered or its impact diminished by modular design and decoupling, including locating a town and residences at a significant distance from the site of the risky activity. The effectiveness of such measures is, however, inversely related to the impact of an adverse event. Locating residences 10 kilometres from a chemical factory would have dramatically reduced the impact of the chemical explosion in Bhopal India in 1984. The explosion and release of deadly chemicals at a Union Carbide pesticide plant by some estimates killed more than 3,700

24 See European Commission, Risk Assessment and Mapping Guidelines for Disaster Management, 2010, pp. 15-16.

people.[25] But in the case of the 1986 reactor meltdown at Chernobyl, people were affected who lived at much greater distance from the plant, and a large region in the Ukraine had to be abandoned due to radioactive contamination.

Most early LAWS will have a relatively low likelihood of causing an adverse event. However, as systems become more autonomous, learn, and can alter their own programming, the likelihood of unpredictable behaviour goes up, as does the likelihood of an adverse event. Some of this unpredictability might be mitigated through software safety techniques, resilience engineering, advances in the development of control algorithms, and additional new methods for testing, monitoring, and managing complex adaptive systems. On the one hand, the research directed at such problems can be anticipated to yield results. On the other hand, presuming that new engineering approaches will be satisfactory in advance of their development is a simplistic form of techno-solutionism and is foolhardy.

Acceptable levels of risk

Member states may differ on the degree of unpredictability, uncertainty, and risk they will accept in the behaviour of the weapons systems they deploy. However, from the perspective of IHL, should the international community bow to such subjectivity?

At the very least, there should be a prohibition on introducing autonomous features such as targeting and firing into platforms launching high-powered munitions. For such weapons systems, the risks of a catastrophic event are high, even while the likelihood of such an event might be relatively low. Setting a payload threshold for a prohibition on autonomous features being introduced into seriously destructive weaponry will be somewhat arbitrary, but nevertheless necessary.

From the perspective of predictability alone, AWS with low firing power are more problematic. Clearly some strategic planners have concluded that the advantages will significantly outweigh any risks. Given that most of the unpredictable actions of such systems will not cause harm to people, non-human animals, and other entities worthy of moral consideration, such estimations are understandable, although strategy analysts frequently under-

25 There have been disputes over the death toll of the 1984 explosion at the Union Carbide plant in Bhopal. All parties agree that the immediate death toll exceeded 2,200. The government of Madhya Pradesh claims a total of 3,787 deaths due to the release of methyl isocynate and other chemicals.

estimate the probability of an unacceptable tragedy. In October 2007, a semiautonomous robotic cannon deployed by the South African army malfunctioned, killing nine soldiers and wounding 14 others.[26]

In an unstable geo-political context, even the unintended use of a weapon with low firepower can have far-reaching consequences, particularly when non-combatants die. How should member states respond when a complex adaptive autonomous system's action leads to civilian deaths, starts a war, or heightens existing hostilities, particularly when such a result differs from the intention of those who deployed the system? Will or will this not be a violation of IHL? What if the harm to civilians involves the deaths of thousands, or hundreds of thousands?

Member states may not concur as to whether a degree of unpredictability alone justifies a prohibition on autonomous targeting and firing features being introduced into systems whose destructive power is low. Nonetheless, in combination with ethical concerns not covered in this paper, arguments for prohibiting low-powered munitions warrant serious attention.

False assumptions

Military strategists from advanced industrial nations propose to policy makers and the public that they will develop autonomous weaponry in a responsible manner and that all systems will be fully tested before they are deployed.[27] Regardless of any belief that such conjectures are made in good faith, they are flawed. First, as noted above, adequate testing of autonomous systems will be difficult, costly, and nearly impossible. Furthermore, even when nations with substantial resources make every effort to ensure the reliability and safety of autonomous systems, the pressures of warfare are such that these efforts will often be truncated.

More importantly, LAWS will not just be deployed by wealthier states capable of developing detailed specifications, designing rigorous protocols for testing, debugging faulty systems, and providing adequate training during the deployment of new weaponry. Poor

26 N. Shachtman, Robot Cannon Kills 9, Wounds 14, Wired, 18 October 2007, http://www.wired.com/2007/10/robot-cannon-ki/.

27 See, for example, the U.S. Department of Defense Directive 2000.09, Autonomy in Weapon Systems. The Directive is dated 21 November 2012 and signed by Deputy Secretary of Defense, Ashton B. Carter, who was appointed Secretary of Defense by President Obama on 5 December 2014; http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf.

nations and aggressive non-state actors, unable or unwilling to carry out all of these activities, will also utilise LAWS in defence of their people and territory, and in pursuit of their goals. In addition, arms manufacturers will assemble many low-cost LAWS by combining readily available software and hardware, initially developed for domestic applications, together with equally available munitions. In other words, assurances as to their intentions by military leaders from major powers will have little effect on the proliferation, reliability, and availability of AWS. No means exist to ensure that other countries, friend or foe, will institute quality engineering standards and testing procedures before they deploy lethal autonomous weapons.

Conclusions

Even crudely intelligent autonomous machines signal an inflection point in human history that will alter the course in which all future wars will be fought. The long-term consequences of the decisions the CCW makes in regards to AWS could far outweigh the short-term benefits.

A rephrasing of the Collingridge dilemma[28] captures an important aspect of the CCW's deliberations. It will be easiest to shape and control the development of LAWS early on. By the time the undesirable consequences are fully realised, LAWS will be so much a part of the economic and social fabric that its control and elimination from the conduct of warfare is extremely difficult, if not impossible.

While we can only perceive the future through a glass darkly, expert testimony at these three years of CCW meetings has attempted to outline the ramifications of relying on AWS for the conduct of future wars. At the very least, the experts have provided a degree of foresight, and an opportunity for member states to take responsibility for acting or not acting on the call to prohibit LAWS. A prohibition would not entirely stop the development of the technology. It would, however, require meaningful human control.

28 The Collingridge dilemma has had considerable influence on technology policy since David Collingridge first proposed the quandary in 1980. In his book "Social Control of Technology", he states that it is easiest to control development early on, "by the time undesirable consequences are discovered (...) the technology is often so much a part of the whole economic and social fabric that its control is extremely difficult"; D. Collingridge, Social Control of Technology, London 1980, p. 11.

Some categories of LAWS certainly deserve an outright prohibition, for example, an AWS that functions as a platform for a nuclear weapon or other high-powered munition (as designated by CCW), or an AWS which can alter its own programming. If, or when, adequate control mechanisms, refined powers of discrimination, sensitivity to morally significant considerations, or tools for predicting system behaviour reliably for AWS are eventually developed, the CCW can revisit the acceptability of specific wartime applications.

The stakes and pressures of succeeding in warfare compel adversaries to deploy weapons systems as soon as they are developed and prove to be adequate for the goals of states and non-state actors. The burden on stopping or even slowing the deployment of a weapons system lies with those who wish to prohibit it. The debate is not being conducted on a level playing field, in that those wishing to arrest any prohibition can use obfuscation, ambiguities, and marginal cases to thwart action by member states and by CCW as a whole.

Given the breadth of systems encompassed and far-reaching consequences of approving AWS, the standard should be reversed as to the acceptability of lethal autonomous weapons. In principle, this overall category of systems should be prohibited, with the proviso that states can petition the CCW to approve certain sub-categories of AWS. For example, nearly all states would agree that defensive systems deployed to intercept incoming missiles are an acceptable application. This exception would be included in any prohibition. AWS used for perimeter control to protect conclaves of refugee non-combatants might well be approved. Perhaps the use of AWS in tightly proscribed theatres of war will find approval. There may be situations in which LAWS could function as a reliable deterrent quelling renewed hostility between known adversaries. The burden as to whether the behaviour of a specific form of LAWS is adequately predictable should be upon those who wish to deploy the weaponry. The proof that such weapons can be deployed safely and with meaningful human control must reside with those parties who petition to use a specific form of AWS.

AWS pose unique challenges for arms control negotiators. Definitional ambiguities are one challenge. In addition, AWS do not fit into traditional regimes for oversight and inspections. Establishing whether meaningful human control was enacted during a specific incident will be extremely difficult. Nevertheless, the stakes are high, so the CCW will need to be creative in forging a way to stop or limit the use of the riskiest AWS.

Researchers in artificial intelligence and robotics already appreciate that the development of truly beneficial and robust autonomous systems is a delicate process requiring care and attention in order to ensure that the systems will be safe and controllable. The scope of the control problem is just beginning to be outlined. Whether the control problem for

autonomous systems can be adequately solved remains unclear. The benefits for humanity are tremendous, thus the pursuit of beneficial AI is a worthy pursuit. However, nothing will imbalance the responsible development of AI more than an unfettered arms race in autonomous systems. It behoves the member states to give researchers an opportunity to develop autonomous systems responsibly.

The degree of predictability in the behaviour of autonomous weapons systems will vary, and will even vary among individual units of a similar design. If the CCW should leave the window to deploying AWS wide open, there will be no criteria for determining which systems do and which do not have a high degree of predictability.

The future will not forgive us if we choose unwisely. And our successors will certainly condemn us if we use definitional ambiguities as an excuse to not act at all in taking a clear position on whether to embrace or prohibit lethal autonomous weapons. The inflection point, the window of opportunity, to shape the development of LAWS is presently open. That window, however, will close quickly once belligerents begin deploying LAWS.