SEARCHING FOR MEANINGFUL HUMAN CONTROL

The April 2018 meeting on Lethal Autonomous Weapons Systems

Barbara Rosen Jacobson

Summary

• This briefing paper analyses the discussions held during the second meeting of the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS).

• The meeting built on the conclusions and recommendations of the November 2017 session, where states agreed on the applicability of international humanitarian law (IHL) and the responsibility of states for the deployment of LAWS.

• Addressing remaining issues of contention, the meeting attempted to provide a deeper understanding of the characteristics of LAWS, as well as the necessary degree of meaningful human control in their development and use.

• There seems to be a growing consensus about the necessity of meaningful human control in the critical functions of LAWS, i.e., selecting and engaging a target, although the concept of ‘meaningful’ remains undefined.

• There is a need for accountability throughout the life cycle of an autonomous weapon, from its development to its use, although there is still a lack of clarity on the distinct responsibilities of different actors involved in the development and use of LAWS.

• Several different policy options were discussed – strengthening Article 36 of Additional Protocol I to the Geneva Conventions, issuing a political declaration, or establishing a legally binding instrument – and while delegates did not agree on a preferred mechanism, there was a growing sense that the policy options are not necessarily mutually exclusive.

• The GGE managed to allow for a deeper understanding of the potential risks (and benefits) of LAWS and there was some convergence of views on concepts such as meaningful human control. Yet, many issues of divergence remain, such as the scope of a definition or the need for a pre-emptive ban – which will have to be addressed in the August 2018 meeting, which is expected to result in a set of recommendations.
In the context of the Convention on Certain Conventional Weapons (CCW), states have been discussing the question of whether policy options are needed to govern the development and use of potential lethal autonomous weapons systems (LAWS). Since November 2017, discussions have taken place in the form of a Group of Governmental Experts (GGE), which met for the second time in April 2018. This briefing paper summarises the main points of debate that arose in this meeting and puts them in the larger context of artificial intelligence (AI) and autonomous decision making.

The first meeting of the GGE in 2017, demonstrated that there is a wide range of positions related to LAWS among high contracting parties (countries that have signed the CCW), ranging from different views on its definition to the desirability of policy options and potential legally binding instruments. Nevertheless, the group was able to agree on a set of conclusions, captured in the Report of the 2017 GGE on LAWS. These include:

1. International humanitarian law (IHL) continues to apply fully to all weapons systems, including the potential development and use of LAWS.

2. Responsibility for the deployment of any weapons system in armed conflict remains with states, which must ensure accountability for lethal action by any weapon system used by the state’s forces in armed conflict.

3. Given the dual nature of technology, the work of LAWS should not hamper progress in or access to civilian research and development and use of these technologies.

4. Given the pace of technological development and uncertainty regarding the pathways for the emergence of increased autonomy, there would be a need to keep potential military applications or related technologies under review.

Based on the 2017 meeting, it was clear that there were a couple of elements that had to be discussed in further detail, especially considering that the discussions often stalled due to a lack of a common definition of LAWS. Therefore, the April 2018 meeting focused on the characterisation of LAWS, as well as considerations of the human element in the use of lethal force, which many see as the defining element of autonomous weapons.

The topic was discussed in much greater depth and breadth than during the November 2017 meeting, with some states calling to simplify and focus the discussion, and others appreciating the depth of the conversation. After five days, it had become clear that despite certain remaining differences, areas of consensus were growing around several issues, most notably the necessity of meaningful human control throughout the development and use of autonomous weapons.

Autonomy and its ethical, legal, and military repercussions

The group’s discussions on LAWS are motivated by the potential ethical, legal, military, and technical challenges they pose, and their possible adverse effects on international peace and security. In this context, some have warned that LAWS could lead to an arms race. They could lower the threshold to the use of force – as it is assumed that they minimise harm to troops – leading to proliferation, and they may even fundamentally change the nature of warfare. In situations where some states acquire LAWS, while others continue to rely on more traditional means of warfare, dangerous asymmetries may arise. There is also a risk that LAWS could be acquired by non-state actors and terrorist groups.

At the technical level, there are risks related to errors in the machine and bias in the algorithms, and the possibility that they could be hacked and interfered with. Several delegations explicitly mentioned the challenge of gender bias that could be incorporated in autonomous systems, especially considering the fact that the field is currently male-dominated (also evident in the general lack of gender balance of the delegations at the GGE itself). Not only are such systems susceptible to ‘technical hacking’, if they rely on perception-based algorithms (such as facial or voice recognition), deliberate changes in the environment in which the systems operate could also contribute to malfunctioning. The algorithm could misinterpret visions or sounds.

Aside from the security questions, there is also the inherently moral question of the ethics of allowing machines to make a decision on taking the life of a human being, which is universally deemed problematic. Autonomous systems neither have the ability for ethical reasoning, which is inherent to life-and-death decisions, nor do they feel guilt or face trauma after taking another person’s life. If such emotions are removed and such decisions transferred to machines, the threshold to attack might be significantly reduced.
Legal questions arise as to whether LAWS could ever be in compliance with international law, in particular IHL and human rights law. When outsourcing a life-and-death decision to a machine, how does this affect the human rights to life and human dignity? Can machines ever comply with the principles of IHL, such as making a distinction between civilians and soldiers, and acting with precaution and proportionality?

While these risks pose significant challenges in their own right, there is a sense that the clock is ticking, generating the need to keep pace with the rapid advancements of innovation.

The other side of the coin: Could autonomous weapons (ever) be beneficial?

Despite these daunting challenges, some states point out that LAWS could have military, and even humanitarian, advantages. Weapons systems with high degrees of autonomy might perform with greater accuracy than systems operated by humans. Throughout the week, there was a growing recognition that LAWS could, in fact, be beneficial for IHL. One example of an autonomous system with humanitarian benefits was provided by the US delegation, presenting the C-RAM system, which intercepts incoming rockets and shells, protecting soldiers and civilians, based on a mix of human decision making and automation. It must be noted that in this particular example, the mission commander needs to decide on launching the system once an incoming rocket has been detected – a decision which they have about 20 seconds to make – and the commander can stop the engagement of the system at any time. While the example was generally considered useful, some questioned the relevance of the example (is it autonomous enough to be relevant for the GGE?), the timeframe of decision making by the commander, the training that is required to operate such a system, and the reliability of the system.

The elimination of human error was often mentioned by those underpinning the potential humanitarian benefits of AI, one expert claiming that human error accounts for about 10% of losses of human life in military operations. Humans act with a half-second delay between stimulus and response. If, for example, a commander decides on engaging a target, and a child suddenly walks out of a front door, the soldier is likely to shoot anyway, unable to respond on time. An autonomous weapons system might be able to make such decisions far faster than humans, and take into account rapid changes in the environment.

This discussion was summarised in one crucial question: If the use of autonomous systems would indeed lower error levels, is it preferable to have a fully autonomous system able to minimise harm to civilians, yet without human control, or is it preferable for humans to continue to make the ultimate decisions on life and death, with the risk of human error?

Finally, the dual nature of autonomous systems was often recalled, where the blurred lines between the civil and military spheres give rise to additional complexities. A restrictive legally binding agreement risks stifling technological innovation in the civilian sphere. At the same time, civilian technologies could potentially be ‘upgraded’ to autonomous weapons systems. In addition, the military might increasingly rely on technology companies for the development of new weapons systems, due to a lack of in-house capacity in the military. In fact, some of these companies already seem to be working with the US military on the development of increasingly innovative weapons systems, raising the question of the responsibility and accountability of private companies involved in the possible development of LAWS.

Searching for a definition

As became evident during the November 2017 meeting, the formulation of a definition of LAWS is one of the most important issues of the debate. In fact, opinions even diverged on the question of whether a working definition is necessary in the first place. Some preferred to keep the debate flexible to be able to further our understanding of the full scope of the topic, while others emphasised the need for a definition to be able to move towards policy options. To move the meeting forward, the Chair proposed four ways in which a definition could be formulated:

- Separative approach: Eliminate characteristics and concepts not relevant to the CCW while gathering those that are definitely relevant.
• Cumulative approach: Add characteristics to a master list and then decide on their relevance.

• Accountability-oriented approach: Define LAWS according to the level of autonomy or loss of human control, and the type of actions handed over to machines.

• Results-oriented approach: Define LAWS according to the consequences that are to be avoided and then work backwards.

The first two approaches tend to be more of a technical nature, defining the technical characteristics that should, or should not, be part of LAWS. Some states preferred a narrow definition, focusing on lethality, total autonomy, self-learning, and with the impossibility of terminating the system once activated – i.e., the lack of a stop-button. They see LAWS as systems that are able to carry out military objectives in a complex environment, without a framework of set rules, systems that are able to identify and modify targets without human control. These states pointed out that there are already systems in use with elements of autonomy, often used defensively, which have not raised issues of non-compliance with international law, even though some of their functions are autonomous. One delegate pointed out that some feel uncomfortable with ‘putting legitimate weapons in the same basket as terminators’.

Some delegations suggested only focusing on weapons with self-learning mechanisms. If LAWS are able to learn and adapt, they will cease to be predictable and reliable. Others encouraged including both ‘dumb’ and ‘intelligent’ autonomous weapons systems, as simple systems can still be highly intelligent in their critical functions, such as machine guns triggered by sensors. At the same time, highly intelligent systems might still be predictable as they will self-optimise.

Fully autonomous weapons – with no human involvement in the critical functions of these systems – in all likelihood do not yet exist. In fact, many claim that they might never be developed, as it is not in the interests of the military to develop a system that is unpredictable, one they cannot fully control. As a result, some felt uneasy with the ‘speculative character’ of the discussion due to the absence of existing LAWS, and noted that it might be difficult to define a weapon that does not exist today.

Most states, however, adopted a broader understanding of emerging technologies in the framework of LAWS, not least as it might be difficult in practice to determine when a system moves from being ‘automated’ to ‘autonomous’ or ‘fully autonomous’. As fully autonomous weapons systems might never be developed in the first place due to their potential military undesirability, these states cautioned against only considering ‘the tip of the iceberg’ and instead suggested including relevant systems with varying degrees of autonomy in the debate, thus avoiding the debate becoming solely concerned with ‘science fiction’. The criteria of only including weapons without a ‘stop button’ or with self-learning capabilities are seen as rather arbitrary preconditions.

The approach that seemed to have gathered the most support is to characterise LAWS as weapons systems with autonomy in their critical functions, i.e., selecting and attacking targets without human intervention. The advantage of opting for such an approach is that it is not dependent on the technical nature of the weapon; it does not matter which computing methods are used to create autonomy, what counts is the nature of human control in the critical functions. Even machines that are not fully autonomous, but perform relevant functions in the targeting cycle autonomously, might be taken into consideration according to this definition. Some pointed out, however, that there needs to be a better understanding of the term ‘critical functions’.

A growing number of states took issue with the word ‘lethal’ in the acronym LAWS. While lethality is officially part of the mandate of the group, there is an increased awareness that weapons not intended to be lethal can nevertheless use force and inflict excessive harm, and should also be included in the debate, as it is the use of force that triggers obligations under IHL, and not the intended lethality of a weapon. At the same time, an absence of the term lethality would make it difficult to differentiate between systems that could theoretically inflict harm – such as autonomous vehicles – and weapons intended to kill human beings. The Chair decided to postpone the discussion on lethality to a later stage, although it is evident that it will eventually need to be addressed, given the many interventions that questioned the utility of including the term.
The most significant step forward during the April 2018 meeting was the awareness that meaningful human control is essential in the development and use of LAWS. The final decision on life or death should always be human. There was agreement that weapons that can change their mode of engagement without human input are dangerous, especially referring to the potential self-learning capabilities of autonomous weapons systems. As a result, states agreed that autonomous weapons systems should be programmed within parameters that cannot be altered by the weapons system itself. Fully autonomous weapons systems that can change the goal function or alter pre-programmed conditions or parameters do not comply with international law, as it makes such systems unreliable and unpredictable, regardless of their sophistication. Human control should be defined in legal principles, and not be seen as 'goodwill'. Some delegates referred to the Martens clause as providing a legal basis for discussions on meaningful human control.

Finding meaningful human control

The Martens clause appears in a number of instruments of IHL, including the preambles of The Hague Conventions, and in the Geneva Convention and its additional protocols. The clause, as it appears in the 1977 Additional Protocol I, reads:

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

While the clause has been interpreted in different ways, some argue that if a means of war is not explicitly regulated through IHL, it can still be considered unlawful if it runs counter to the 'principles of humanity' or the 'dictates of public conscience'.

What is the meaning of meaningful?

While there was a consensus on the need for meaningful human control, it remains difficult to determine what constitutes meaningful or control. The discussion was enriched by a framework provided by the Chair, which divided the lifecycle of a weapon into four phases:

- Research and development
- Testing and evaluation, verification and validation, and weapons review
- Deployment, command, and control
- Usage and abortion

The delegates and experts at the meeting addressed all phases, although some placed greater importance on the need for human involvement in some of the phases than in others.

There was significant discussion on the role of engineers in developing the systems and their ability to code humanitarian principles into weapons, with notable scepticism.
about the degree to which this was deemed possible at all. Some experts pointed out that there is surprisingly little knowledge of the technologies underpinning current autonomous systems, including autonomous vehicles, and there is a lack of awareness of how to verify and validate such technologies, with one of them stating: ‘We have no idea what we’re doing when it comes to certification.’ In addition, some called for the need to include ethics and law in the curricula of AI systems engineers, as well as for training in science and technology for lawyers and policymakers.

The interaction of the commander with the machine was also subject to discussion. Does a generic command suffice, or should the commander be able to engage, intervene, and override machine functions? In addition, some pointed at a certain delay that could be built into the system, allowing commanders and operators of the system to make decisions. Without such safeguards, machines might respond to sensory input too quickly for commanders to be able to intervene, even if the possibility of intervention is theoretically part of the system.

Moreover, the human operator should have sufficient information and should be able to make an informed decision about whether to follow the machine’s recommended course of action. A proper understanding of the technological aspects of autonomous systems is required to be able to judge the decisions made by the machine. Without sufficient knowledge, machines can influence human decision making, distorting the perception of the commander. This can be especially problematic when the algorithms at the basis of the machine reflect bias and errors.

The feasibility of a one-size-fits-all solution

The issue of human control over LAWS is further complicated as their impact is highly dependent on the context in which they would be deployed. LAWS deployed in maritime environments will have a different effect on their surroundings than those activated in urban warfare. Their effects will be different when they are used offensively or defensively, for a long or short period of time, and whether they are stationary or mobile. Some claim that the influence of the context is so significant that there can be no one-size-fits-all standard.

Human control for human accountability

There was widespread agreement that the human person is the very foundation of the law. Ethical decisions cannot be taken by algorithms. As stipulated in the summary of the November 2017 meeting, states remain responsible for the development and use of LAWS, and must ensure accountability and criminal liability.

For systems that act on the basis of machine-learning, as opposed to rule-based, pre-programmed algorithms, it might be particularly difficult to identify accountability and legal redress, as the decisions made by the system cannot be explained and are developed in a ‘black box’. There may be an accountability gap in a victim’s right to remedy. In this regard, simpler, rule-based AI systems, that are explainable and transparent in their decision making might be preferable. Nevertheless, fully autonomous systems might perform better than such rule-based systems in eliminating human error, and they are less likely to be hacked.

It became clear that the need for accountability and human responsibility should be present throughout the lifecycle of a weapon, from its development to its use. Human involvement and accountability starts with the creation of the algorithm and the programming of the weapon and continues during the phase in which the weapon is tested and reviewed. States should be responsible for properly training those who interact with the system and issue clear guidelines. Those who are taking decisions with regard to the development and use of the systems should be aware of their shortcomings and level of error. Political decisions on the procurement of such weapons will have to be scrutinised, as well as the decisions of commanders to deploy the systems. Military personnel will have to monitor and review its functioning.

While the breakdown of accountability in the different phases of the weapon’s lifecycle provided fruitful insights, some questioned whether this distributed accountability could be a risk in itself. If everyone is considered accountable in every phase, who bears the ultimate responsibility for the weapons system if it is misused or it malfunctions? Most seemed to agree that this responsibility would remain with the commander who decides to launch the system. To make matters more concrete, some suggested learning from other autonomous systems to identify a liability chain, such as autonomous vehicles.
Given the many concerns that have arisen in relation to LAWS, as well as the agreed need for meaningful human control over such systems, the central question related to the policy options of the GGE is whether existing international laws are comprehensive enough to address the risks. And if they are not, do we need to improve the implementation of current laws, or create new legal instruments?

**Strengthening Article 36**

Not everyone agreed that new regulation is needed. Even if fully autonomous weapons turn out not to comply with IHL, this will be detected through mandatory weapon reviews under Article 36 of Additional Protocol I to the Geneva Conventions, which should ensure that no unlawful weapon will be developed and used. Some are of the opinion that IHL already provides a coherent system of regulation.

At the same time, many delegates noted that the implementation of Article 36 should be strengthened if it is to effectively address the concerns of LAWS. In evaluating autonomous systems, many believe that a criterion of meaningful human control should be part of the assessment. Many claimed that there is a need to agree on standardised, universal mechanisms, since they are currently conducted by states themselves and are limited in their oversight, comparability, and harmonisation. In this area, there were many calls for greater transparency, standardisation, confidence-building measures, and the sharing of experiences and best practices. In addition, there might be a need to establish an independent observatory to closely follow developments in the creation and use of autonomous weapons systems and their possible repercussions.

Another concern relates to the feasibility of the application of Article 36 in evaluating self-learning or self-programmable machines in particular, as their characteristics might transform each time such systems are used. Will the review need to be conducted after each operation to ensure the system still complies with IHL?

**Issuing a political declaration**

A Franco-German proposal for a political declaration is gaining increasing momentum. The envisioned declaration should affirm that state parties share the conviction that humans should continue to make the ultimate decision on the use of lethal force and exercise sufficient control over LAWS. The proposal was considered by many to be a fruitful middle ground for the near future, striking a balance between the interests of all parties. Proponents of the declaration argued that after a potential political declaration is developed, states might be in a better position to define policy options, including a possible code of conduct or legally binding instrument.

**Negotiating a legally binding instrument**

A growing number of states were in favour of a legally binding instrument prohibiting the development and use of LAWS, and a moratorium on their current development and use, based on the ethical, legal, military, and technological risks they pose to humanity. They claimed that current IHL and the Article 36 review mechanism does not provide adequate protection against the concerns generated by potential LAWS; ‘If everything would be resolved by Article 36, we wouldn’t have had additional protocols.’ A new norm is needed to provide international legal clarity about the extent of autonomy permissible in weapons systems. Regardless of whether LAWS currently exist or are intended to be developed, they may need to be pre-emptively banned. As one delegate mentioned: ‘If history is any teacher, anything that is even remotely possible becomes a reality.’ A ban on a system that does not yet exist is not even unprecedented, and many referred to the Protocol on Blinding Laser Weapons in the framework of the CCW, which was issued before such weapons had even been developed.

Given the pace of innovation, many states hoped to be able to start negotiations on an additional protocol to the CCW on LAWS towards the end of 2018, after a recommendation of the upcoming meeting of the GGE in August 2018.

Those not in favour of adopting such an instrument claimed that there is currently a lack of proper understanding of the exact risks and potential benefits of autonomous systems in warfare. A pre-emptive ban could have unintended consequences; views may change over time, as we find new ways to benefit from technology. In addition, there exists a risk that such instruments would limit innovation in civilian applications of autonomous technologies.
Many delegations noted that they were pleased with the progress in better understanding the challenges posed by LAWS and with the awareness on the necessity of a minimum level of meaningful human control in the critical functions of weapons systems. In addition, the policy options that were presented, from strengthening review mechanisms to a political declaration or legally binding instrument, were not seen as mutually exclusive, and there was some agreement to at least address the issues related to Article 36 weapon reviews.

Moreover, the number of states that are in favour of a ban on fully autonomous weapons has grown, with three countries joining the list of those in favour of a prohibition on the development and use of such systems. In addition, one country emphasised that it supports a legally binding prohibition of the use of LAWS, but not of their development.

While this meeting allowed for a broad and complex discussion, preparing the grounds for the August 2018 meeting, the debate is expected to be narrowed down for the upcoming meeting, and issues of divergence will have to be addressed in order to formulate recommendations.
Endnotes


Working papers submitted for the April 2018 meeting


Ms Barbara Rosen Jacobson is Programme Manager at DiploFoundation (Diplo) and the Geneva Internet Platform (GIP), where she manages diplomacy and Internet governance-related projects, including Diplo’s activities in the area of Data Diplomacy and GIP’s Data Track. Before joining Diplo in 2014, she worked on several human rights and development projects in Geneva, Indonesia, and the Occupied Palestinian Territories. Ms Rosen Jacobson holds a BA in Social Sciences from Amsterdam University College and an MA in International Affairs from the Graduate Institute of International and Development Studies in Geneva.

We look forward to your comments – please e-mail them to barbarar@diplomacy.edu

Diplo’s policy papers and briefs can be downloaded from www.diplomacy.edu/policybriefs

If you are interested in publishing a policy paper or brief with us, please get in touch with Katharina Höne, at katharinah@diplomacy.edu