EXPERT MEETING

AUTONOMOUS WEAPON SYSTEMS

IMPLICATIONS OF INCREASING AUTONOMY IN THE CRITICAL FUNCTIONS OF WEAPONS

VERSOIX, SWITZERLAND
15-16 MARCH 2016

ICRC

REPORT

ICRC

International Committee of the Red Cross
19, avenue de la Paix
1202 Geneva, Switzerland
T +41 22 734 60 01 F +41 22 733 20 57
E-mail: shop@icrc.org www.icrc.org

EXPERT MEETING

AUTONOMOUS WEAPON SYSTEMS

IMPLICATIONS OF INCREASING AUTONOMY IN THE CRITICAL FUNCTIONS OF WEAPONS

VERSOIX, SWITZERLAND
15-16 MARCH 2016

CONTENTS

INTRODUCTION AND STRUCTURE OF THE REPORT

Debates on autonomous weapon systems have expanded significantly in recent years in diplomatic, military, scientific, academic and public forums. In March 2014, the ICRC convened an international expert meeting to consider the relevant technical, military, legal and humanitarian issues.[1] Expert discussions at a Meeting of Experts convened by the High Contracting Parties to the UN Convention on Certain Conventional Weapons (CCW) were held in April 2014 and continued in April 2015 and April 2016.[2]

As a further contribution to the international discussions, the ICRC convened this second expert meeting, entitled *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, from 15 to 16 March 2016. It brought together representatives from 20 States[3] and 14 individual experts in robotics, law, policy and ethics.

This report of the meeting is divided into three main sections:

Part I is a summary report of the expert meeting, which was prepared by the ICRC under its sole responsibility.

Part II comprises summaries of selected presentations given by individual experts at the meeting, and provided under their own responsibility.

Part III is an edited version of the background paper prepared by the ICRC and circulated to participants in advance of the expert meeting in March 2016.

The meeting programme and the list of participants are provided in **Annexes 1** and **2**.

[1] ICRC (2014) Autonomous weapon systems: technical, military, legal and humanitarian aspects, https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf.

[2] CCW Meetings of Experts on Lethal Autonomous Weapon Systems (LAWS), 2014, 2015 and 2016,. http://www.unog.ch/80256EDD006B9C2E/(httpNewsByYear_en)/0462FC37E62E7E73C1257E2A005A013A?OpenDocument

[3] Algeria, Australia, Brazil, China, Egypt, France, Germany, India, Israel, Japan, Mexico, the Netherlands, Pakistan, the Republic of Korea, the Russian Federation, South Africa, Sweden, Switzerland, the United Kingdom and the United States.

PART I: SUMMARY REPORT PREPARED BY THE INTERNATIONAL COMMITTEE OF THE RED CROSS

Expert Meeting on Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons, 15–16 March 2016, Versoix, Switzerland.

A. BACKGROUND

Debates on autonomous weapon systems have expanded significantly in recent years in diplomatic, military, scientific, academic and public forums. In March 2014, the ICRC convened an international expert meeting to consider the relevant technical, military, legal and humanitarian issues.[1] Expert discussions within the framework of the UN Convention on Certain Conventional Weapons (CCW) were held in April 2014 and continued in April 2015 and April 2016.[2]

Discussions among government experts have indicated broad agreement that "meaningful", "appropriate" or "effective" human control over weapon systems and the use of force must be retained, but there has been less clarity on the type and degree of control necessary from a legal, ethical and policy perspective. The ICRC has called on States to set limits on autonomy in weapon systems to ensure that they can be used in accordance with international humanitarian law (IHL) and within the bounds of what is acceptable under the principles of humanity and the dictates of the public conscience.[3]

In view of the incremental increase of autonomy in weapon systems, specifically in the "critical functions" of selecting and attacking targets, the ICRC has stressed that experience with existing weapon systems can provide insights into where the limits on autonomy in weapon systems should be placed, and the kind and degree of human control that is necessary to ensure compliance with IHL and ethical acceptability.

With this in mind, the ICRC held its second expert meeting, entitled "Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons", from 15 to 16 March 2016. It brought together representatives from 20 States[4] and 14 individual experts in robotics, law, policy and ethics, and was held under the Chatham House Rule.[5] The six sessions reflected the overall objectives of the meeting, which were to:

- consider the defining characteristics of autonomous weapon systems;
- better understand autonomy in the critical functions of existing weapon systems;
- explore emerging technology and the implications for future autonomous weapon systems;
- examine the legal and ethical implications of increasing autonomy in weapon systems;
- consider the legal, military (operational) and ethical requirements for human control over weapon systems and the use of force; and
- share approaches to addressing the challenges raised by increasing autonomy.

[1] ICRC (2014) Autonomous weapon systems: technical, military, legal and humanitarian aspects, https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf.
[2] CCW Meetings of Experts on Lethal Autonomous Weapon Systems (LAWS), 2014, 2015 & 2016, https://www.unog.ch/ccw
[3] ICRC (2016) Views of the ICRC on autonomous weapon systems, CCW Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), 11-15 April 2016, Geneva. Background paper, 11 April 2016, https://www.icrc.org/en/download/file/21606/ccw-autonomous-weapons-icrc-april-2016.pdf.
[4] Algeria, Australia, Brazil, China, Egypt, France, Germany, India, Israel, Japan, Mexico, the Netherlands, Pakistan, the Republic of Korea, the Russian Federation, South Africa, Sweden, Switzerland, the United Kingdom, and the United States.
[5] https://www.chathamhouse.org/about/chatham-house-rule.

This summary of the presentations and discussions is provided under the sole responsibility of the ICRC and reflects the key points raised by speakers and participants at the meeting.[6]

B. SUMMARY OF PRESENTATIONS AND DISCUSSIONS

1. Characteristics of autonomous weapon systems

Speakers in this session debated the defining characteristics of autonomous weapon systems with a view to clarifying the terminology and fostering a better understanding of the types of weapons under consideration. The ICRC's working definition was used as a basis for discussions throughout the meeting, although at times some speakers and participants expressed a different understanding of definitions. Under the ICRC's definition, an autonomous weapon system is:

Any 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.

In explaining the working definition, the ICRC emphasized that it was not being used as a means of normative development or to establish a prohibition. Rather, it enabled consideration of the full range of relevant weapon systems, including existing weapons with autonomy in their critical functions that do not necessarily raise legal issues. The ICRC explained that the definition is based on the role of the human rather than the "degree of autonomy", and encompasses any weapon that could independently select and attack targets, whether described as "highly automated" or "fully autonomous". The rationale for that approach being that all such weapons raise the same core legal and ethical questions: in the intended circumstances of use, can the weapon system select and attack targets in a way that respects the rules of IHL? In cases where operation of the weapon system results in an apparent violation of IHL, would it be possible to attribute responsibility to an individual or a State, and to hold them accountable? Is it ethically acceptable (based on the principles of humanity and the dictates of the public conscience) for the weapon system to independently select and attack targets?

One speaker explained that there was no difference, from a technical perspective, between an "automated" and an "autonomous" system, since they could both operate without human intervention after initial activation. And indeed, all three speakers concurred that there was no clear line between "automated" and "autonomous" weapons. The speaker suggested, therefore, that an autonomous weapon system could be conceived of as one with a high degree of automation in relation to software-controlled "safety- and security-critical systems", i.e. systems that could cause danger, harm or even death if they malfunctioned. However, the speaker noted that the "level" or degree of autonomy of a particular weapon system would also be related to the circumstances in which it was employed.

The speaker explained that any autonomous weapon system would always have a model defining the environment within which it could operate. Any operation outside that environment – or unforeseen changes to the environment – would necessarily lead to unpredictability in its functioning. The speaker added that great caution was needed in any development, testing and deployment of such systems to ensure that they functioned as intended in the environment in which they were designed to be deployed.

[6] Detailed summaries of some presentations are provided, under the responsibility of the speakers, in Section III of this report. Some of these summaries provide supplementary information to that presented during the meeting.

Another speaker warned of projecting human behaviour onto machines, and argued that autonomy in weapon systems should be assessed by looking at which parts of the targeting decision-making process were delegated to the weapon system. The speaker noted that some tasks of the targeting process had been delegated to machines for some time. While noting the importance of "selecting and attacking targets" as critical functions, the speaker argued that the human role in other parts of the targeting process would also influence the legal acceptability of a particular weapon system. The speaker suggested that problems would arise when too many tasks in the overall targeting decision-making "loop" were delegated to machines, as that would be the point at which humans risked delegating the decision to kill.

The speaker said that the key question for compliance with IHL would be predictability (i.e. knowledge of how the machine will function in a given context), arguing that autonomy was limited in existing "highly automated" weapon systems – and predictability maintained – owing to restrictions on the scope of their tasks and their context of use. If it was not possible to reasonably predict that a weapon system would comply with IHL, he added, then it would potentially be "unlawfully autonomous".

During the discussion, there was a debate among participants about whether autonomy should be considered a binary feature or rather a sliding scale. One participant argued that autonomy should be assessed at the level of the complete weapon system, and that autonomy in specific functions would not necessarily make a weapon autonomous. Some participants took the view that discussing autonomy in "selecting and attacking" targets, which would include some existing weapon systems viewed as legal by States, was too broad an approach, in particular for regulation purposes. One participant stressed that a narrower definition would be needed for the purpose of States agreeing to regulation. Another participant supported the ICRC's approach, arguing that starting with a broad definition enabled analysis of existing weapons to assess which specific parameters determined compliance with IHL.

The question of the predictability of the weapon system was discussed with great interest. One participant inquired how predictability could be assessed realistically during testing, and a speaker acknowledged that determining in advance how an autonomous weapon system would operate in real-world environments would raise challenges. Using the example of the battle of Fallujah in 2004, during which US Marines had needed to distinguish between civilians and combatants in a split second, the speaker said that it might never be possible to predict how a machine would handle such a situation, although a machine would not approach the situation in the same way. For example, a machine could perform tasks differently and wait longer than a soldier for indicators before targeting a person. The speaker added that uncertainty about IHL compliance might be addressed through programming restrictions on the scope of the machine's tasks.

Another participant asked how adaptation and machine learning in autonomous weapon systems could be reconciled with predictability, and a speaker said that one had to look at the effect that adaptation had on IHL compliance, which might depend on the specific parameters under which the system could adapt its functioning. For example, a system might be authorized to "learn" and adapt in some functions, while it was strictly limited in others. That could be done through programming, e.g. by allowing the machine to do anything except x, y, and z, or by physical limitations in the hardware, e.g. that prevented the machine from carrying out an undesirable action. The speaker added that undesirable consequences were not necessarily limited to the attack itself. For example, a ground robot that was programmed to target a certain object under strict limitations, but that had complete freedom as to how it navigated to the object, could cause civilian damage en route by driving through a village. Therefore, limits on such behaviour would need to be set at the programming stage.

One speaker stressed that adaptation would certainly raise significant questions about predictability, and therefore questions of compliance with IHL, since not knowing when, how, and where a machine would carry out an attack would prevent the user, or commander, from being able to implement his/her legal obligations with respect to the conduct of hostilities. Another speaker emphasized that, technically speaking, it would be extremely difficult to develop a machine that could adapt its functioning to changing circumstances.

2. Autonomy in existing weapons

The second session of the meeting examined autonomy in the critical functions of existing weapon systems with a view to a better understanding of their functioning and how human control over their operation is implemented.

2.1 *Missile- and rocket-defence weapons*

This sub-session considered missile- and rocket-defence weapon systems, commonly used for short-range defence of ships or ground installations against missiles, rockets, artillery, mortars, aircraft, unmanned systems and high-speed boats.

The first speaker provided an overview of the technical operation and military utility of the Counter-Rocket, Artillery, and Mortar (C-RAM) system, which is used to defend military bases from incoming attacks. The speaker noted that the main drivers for the development of the C-RAM were the need for increased precision and accuracy and fast reaction times for defending against attacks. The system has some autonomy in detecting, tracking, selecting and attacking targets; however, the speaker emphasized that the decision to attack is retained by the commander, who decides when to activate the system in a given circumstance, retains oversight over the system during its operation, and is able to deactivate the weapon to stop an attack at any time. The speaker also noted that the weapon system was periodically reviewed by lawyers to ensure that it could be used lawfully.

The speaker explained that, during operations, the computer command-and-control component of the weapon system is constantly updated with information about commercial (civilian) aircraft flight paths and “friendly” aircraft. Based on that information, the computer determines “engagement zones” within which it will carry out attacks once activated. The speaker added that the system employs self-destructing rounds (bullets) to minimize the risk to civilians or others should the rounds miss their target.

The second speaker discussed similar weapon systems with autonomy in detecting, tracking, selecting and attacking targets, including the Iron Dome and the Terminal High-Altitude Area Defence (THAAD) systems. The Iron Dome is a type of counter-rocket, artillery and mortar weapon system capable of intercepting multiple targets at short range. The speaker noted that the system had been shown to be almost 90% effective at intercepting targets, although there were instances where it had misidentified “friendly” aircraft as potential threats. The THAAD system is used for longer-range defence against missiles, and also operates autonomously; a long-range radar detects and tracks an incoming missile, calculates its trajectory and then attacks it with an interceptor missile.

The speaker also explained that the performance of these autonomous missile- and rocket-defence weapon systems could be influenced by a number of different factors, in particular: the technical configuration of computational units, seeking radars, control algorithms and missile controls; the speed of communication between different components of the system; and the accuracy of targeting systems. The speaker predicted that, in the future, smaller defensive systems might be increasingly used for perimeter security, and also posited that, if weapon systems were to be deployed in outer space, they would likely have a high degree of autonomy due to communication challenges in that environment.

During discussions, there was continued debate about the definition of an autonomous weapon system, and whether the weapons described should be considered "highly automated", "semi-autonomous" or "autonomous". Independent of definitions, one participant said that it would be useful to further examine which aspects of the human-machine interaction in the use of those weapons ensured their compliance with IHL, including restrictions on their operation in time and space, and the measures taken to ensure that only legitimate targets were attacked.

There were also questions raised on whether the speed of operation could realistically allow sufficient time for human intervention, and whether, and how, the described defensive systems permitted assessments of the risks of civilian casualties. To the first question, a speaker responded that the C-RAM weapon system described operated for limited times, and to the latter, that there had not been any "collateral-damage" incidents reported in the past 11 years.

To the question of whether there was a clear distinction between "offensive" and "defensive" weapon systems, one speaker responded that the question would be determined on a case-by-case basis. Another participant pointed out that all the systems discussed during the session were anti-materiel weapons, and therefore would not be considered "lethal" from that participant's point of view.

2.2 *Vehicle "active-protection" weapons and anti-personnel "sentry" weapons*

This sub-session examined two quite different types of weapons with autonomy in selecting and attacking targets: vehicle "active-protection" weapons, which are designed to protect armoured vehicles from attacks with missiles, rockets, and rocket-propelled grenades; and anti-personnel "sentry" weapons, which have been developed for the defence of specific sites, perimeters or borders.

The speaker explained the operation of those weapons using two examples. The Trophy (ASRPO-A) active protection system, which is fitted to tanks and armoured vehicles, is employed to defend against incoming threats, such as rocket-propelled grenades, and has been used operationally for five years. Once activated, it employs a radar to detect threats on an incoming trajectory and, if the computer judges that the incoming munition would hit the vehicle, it autonomously attacks by firing small metal balls.

The speaker went on to discuss an anti-personnel "sentry" weapon called Sentry Tech, which is an automated gun system that can incorporate light weapons and anti-tank weapons. The system, mounted on a pillbox, uses computerized sensors with some degree of autonomy to detect and identify human targets. However, the speaker explained that the decision to select and attack a human target is retained by operators who, following an alert from the computer system, initiate an attack by remote control from a distant control station. While some functions of the targeting process (such as detecting and identifying targets) are delegated to the machine, the action of launching an attack remains a human decision.

During discussions, one participant raised questions of whether the acceptable degree of autonomy in weapon systems depended on the nature of the threat, and whether the system was offensive or defensive. Another participant concurred with the speaker that a "fully autonomous weapon system" would not be desirable, but that the key question was where to draw the line with increasing autonomy. The speaker responded that, in his view, a weapon system was acceptable as long as the commander or operator retained control over the decision to kill. In that respect, the speaker stressed that if no person could be found responsible for the actions of the weapon system, that weapon system would not be acceptable; at least one person must be accountable for the system's operation.

Participants asked whether the use of the Trophy weapon had resulted in any civilian casualties and whether the system was continuously used in autonomous mode. The speaker said that there had not been any civilian casualties reported, and explained that the Trophy was normally activated for the duration of an operation (e.g. the journey of an armoured vehicle), but that it would only launch an attack if it detected an incoming threat. Another participant asked what would happen if the system were to be mounted on an unmanned vehicle and the communication with the weapon system were broken; would the system continue to operate autonomously without human oversight? The speaker responded that the system should not be used if communication channels were unreliable. If the human operator decided to allow the system to operate despite a loss of communications, the operator would be held accountable for that decision. In any case, the speaker said, the commander would be responsible for any use of the system that would result in disproportionate civilian harm.

One participant asked whether the Sentry Tech could also fire autonomously, to which the speaker replied that it was only used to fire by remote control. Another participant noted that the Korean "sentry" weapon systems, referenced in the ICRC's background paper for the meeting, also did not select and attack targets autonomously, but had a human "in the loop" to launch an attack by remote control.

2.3 *Sensor-fused munitions, missiles and loitering munitions*

The speaker in this sub-session focused on autonomy in missiles and loitering munitions. Missiles have on-board guidance systems, and they generally fly to a pre-programmed or designated location. Some missiles then use inbuilt sensors, such as active radar, and information-processing capabilities, such as automatic target recognition software and pre-programmed signatures of target objects, to determine their specific target. Loitering munitions operate in a similar way, but have more freedom to search for, select and attack targets over a designated area and time period, using on-board sensors and pre-programmed target signatures.

The speaker explained that many different variables influenced the level of autonomy in missiles and loitering munitions, which could be divided according to three indices: self-mobility (e.g. the ability to move and navigate autonomously); self-direction (e.g. the ability to identify and discriminate targets autonomously); and self-determination (e.g. the ability to launch an attack or adapt its functioning autonomously, e.g. by setting its own goals or choosing targets).

The speaker emphasized that missile technology was becoming increasingly automated, and more systems were programmed to fly to a location in space. Once in that area, they use active sensors to identify, acquire and fire on a target. Those systems, including missiles and loitering munitions, have higher levels of self-mobility and self-direction. An example given was the Long-Range Anti-Ship Missile (LRASM) (currently in development), which has a high level of autonomy in both mobility and navigation, as well as in detecting, selecting and attacking targets. Once the missile arrives at a location in space, it uses on-board sensors to determine its target. The speaker also explained that there was the potential for even greater autonomy with loitering systems, which are programmed to search over a wider area rather than flying to a specific location. The speaker provided examples of loitering munitions that are currently "human in the loop" for target selection and attack, as the types of weapon system that could become autonomous in the future; for example, the Tactical Advanced Recce Strike (TARES) anti-materiel loitering munition has a 200 km range, a 4-hour flight time and carries a 20 kg warhead, and the Hero 30 anti-personnel loitering munition has a 40 km range, a 30-minute flight time and carries a 0.5 kg warhead.

In the view of the speaker, increasingly autonomous weapons are likely to emerge as novel combinations of existing weapons technology rather than entirely new systems; for example,

unmanned weapon platforms equipped with highly automated submunitions. The speaker argued that the extent to which increasing autonomy would raise challenges in terms of human control over "critical functions" would depend on the task delegated to the weapon (broad or narrow), the amount of planning (or changes to planning) relating to that task, and the capability of the system to discriminate targets. An expansion of the weapon system's freedom of action in time and space, the speaker added, would also have implications for both the predictability and reliability (i.e. knowledge of how often the machine will function as intended) of the weapon.

The speaker also highlighted some emerging technology, such as research being undertaken to design Automatic Target Recognition (ATR) software that would incorporate machine-learning technology, so that new targets could be learned in real time and the on-board target library updated accordingly. The speaker emphasized that achieving that goal would present significant technological challenges.

During discussions, several participants commented that weapon systems with machine-learning capability would raise serious questions about predictability. One speaker and a participant explained that machine-learning systems were, by definition, unpredictable. One participant explained that machine learning is not related to the concept of learning in humans. A machine might "learn" to recognize a specific image, but it only recognizes the image based on what it has "seen" previously. Such machine learning might be done in advance or during the operation of the machine. However, the machine has no understanding, in a human sense, of the nature or concept of that object.

Some participants agreed that there could be no predictability where it was not possible to foresee what a machine would do within the parameters of its programming. One participant said that it was hard to see how a system that could self-learn and adapt its own functioning would pass a legal review, as it would not be predictable; in principle, any such modifications in functioning would require a new legal review, as it would become a new weapon. Another participant added that a system that had the ability to attack a broad range of different military objectives, or could move from one target to another, would also raise questions of predictability and compliance with IHL. One participant said that those assessments might also depend on the specific type of weapon and the particular environment of its use. One of the speakers emphasized that a key question was whether it was acceptable to program a machine to select and attack a very broad class of targets, or whether predictability implied the need to programme specific targets. In other words, the fewer constraints on targeting, the more problems would arise for IHL compliance. The speaker added that it was necessary to look at the inbuilt limits of the machine: could the machine attack several military objectives in a row, without returning to its base? Could it target a wide variety of military objectives or was it limited to a specific type, e.g. tanks?

2.4 *Torpedoes and encapsulated torpedo mines*

The speaker in this sub-session discussed a range of torpedo weapon systems with differing levels of autonomy in selecting and attacking targets. The Sea Hake heavyweight torpedo has a sonar to detect its target after launch, but it is connected via a cable to the operator, and retains a "human in the loop" who can redirect the torpedo. The MU90 lightweight torpedo, on the other hand, is a "fire and forget" weapon which, after launch, uses its own sensors to detect and attack a target submarine, and is programmed not to operate above a certain depth. Another anti-submarine "fire-and-forget" weapon discussed was the SHKVL rocket-propelled torpedo.

The speaker also described the Mark 60 CAPTOR encapsulated torpedo and the PMK-1/2 propelled sea mine. The Mark 60 CAPTOR is tethered to the seabed and uses pre-programmed signatures of submarines to autonomously detect and then attack by launching

a torpedo. Self-propelled sea mines, such as the PMK-1/2, function in a similar way, but the whole weapon system moves to attack the target submarine.

During discussions, one participant asked whether it was possible to communicate with the sea mines, for example, to update the signatures the mine used to identify its target, and whether increased autonomy in sea mines was likely in the future. The speaker explained that the systems described did not allow communication after emplacement; however, to reduce the risk of unintended targeting, States would provide details to other States about where the mines had been placed. The speaker also mentioned, in response to a question about the persistency of mines, that some would shut down after a maximum number of weeks, whereas others would remain active as long as the battery allowed. Regarding future developments, another participant said that a particular increase in autonomy for torpedoes and sea mines was not foreseen.

One participant observed that it could be easier to develop autonomy in a maritime environment, since the environment was less cluttered than in ground warfare. However, another participant said that that was less and less the case, as there were an increasing number of civilian objects in the maritime environment, including vehicles used for scientific and industrial tasks, among other civilian purposes. In any case, it was stressed that a major driver for autonomous undersea systems was the difficulty of communicating in that environment. Another participant emphasized that this inability to communicate could raise concerns, especially for weapon systems which operated over long loiter times without the possibility for human intervention.

One participant asked how it was possible to distinguish military targets from protected objects, such as hospital ships and civilian vessels, and what procedure States observed after a mine was no longer needed. The speaker explained that distinguishing military targets from civilian objects was possible owing to the different acoustic signatures of ships and submarines, and that those weapon systems had well-developed target libraries that would help ensure that civilian ships would not be sunk. The speaker added that, when mines were no longer needed, they may shut down, and, if the terrain allowed, they might be physically removed from the area.

3. Emerging technology and future autonomous weapons

Looking to the future, this session sought to examine emerging technology developments in order to consider the potential nature of future autonomous weapon systems.

The first speaker explained that the level of autonomy of a particular weapon system was related to the level of human intervention in the functioning of the system, i.e. both the degree of human control and the point at which such control was exercised. For example, he explained that existing stationary missile- and rocket-defence systems operate autonomously 95% of the time, but human intervention at specific points during that operation helps ensure that human control is maintained over the use of force.

The speaker emphasized the potential mobility of future autonomous weapon systems as a key characteristic that could lead to loss of predictability and loss of human control in emerging systems. The speaker said that, owing to the increased complexity of the system itself, and the increased complexity and variation in the environment in which it operated, it would be very difficult to predict how mobile autonomous weapon systems would operate. That, in turn, would raise questions about how to test and determine the reliability of such systems. The speaker added that the risks associated with increasing autonomy would also be influenced by the specific task for which the weapon system was used; for example, an autonomous quadcopter (a helicopter propelled by four rotors) fitted only with a camera (and

not weaponized) might be considered an acceptable risk owing to the low probability of harm to civilians from a failure or accident.

Speakers also touched upon the drivers for increased autonomy in the "critical functions" of weapon systems on land, in the air and at sea. In this respect, one speaker said that autonomy might enable: increased mobility of robotic weapon platforms; operations in communications-denied environments; a shortened targeting decision-making "loop"; and increased performance over human remote-controlled systems. Another speaker emphasized the military's need for robotic systems that could operate in complex environments and those in which communications were jammed, as well as its desire to reduce the number of human operators.

The second speaker envisaged that advancements in the field of sensors and computing would enable increasing autonomy in military robotics while also being accompanied by increasingly wide access to the technology. He mentioned that current autonomous weapon systems could only operate in specific narrow situations, but that future systems might be designed to operate in more varied and complex environments. In terms of "machine learning", the speaker emphasized that there was still a lack of understanding of how a machine "learned". He said that a machine would either select an option among a range of programmed options, or would develop its own options based on its programming, adding that greater complexity of the machine, and its programming, also increased its unpredictability.

A third speaker offered some additional observations based on developments in civilian robotics. He said that the overall trend was towards supervised autonomy, since sensors were not able to provide machines with a sufficient understanding of changing environments to allow full autonomy. He explained that developments in machine learning would lead to significant improvement in those capabilities, such as image recognition, in the coming years. However, a major challenge would be the lack of predictability as to how such systems would function in any given environment, which in turn would be accompanied by difficulties in testing the systems to determine their reliability.

The speaker said that it was a misconception that only sophisticated, human-like artificial intelligence would allow machines to take decisions. Decisions to take specific actions could today be delegated to supervised autonomous machines. The speaker added that it was easily conceivable that civilian robotic systems could be modified and adapted as weapon systems.

During the discussion, there was a further question about how to test both the reliability and predictability of autonomous weapon systems. One speaker explained that there were no standards in the civilian field for testing autonomous systems. There was a lack of agreement on how to measure their performance and what level of failure was to be tolerated. Another speaker added that it would be very hard to assess reliability at the level of the whole system, but that it might be easier to assess for a specific function. A participant also raised the prospect of swarms – or self-organizing – weapon systems. Such systems, the participant said, would also raise significant questions of predictability and reliability with increasing autonomy. One of the speakers posited that swarm technology remained very challenging, and there were not yet any real-world applications.

One participant suggested that there might be a convergence of autonomous weapon systems and cyber weapons in the future, since the latter might be used to attack the former. Another noted that legal reviews would need to consider autonomous weapons at the system level, assessing both weapon platforms and the specific weapon controlled by the platform.

4. Legal and ethical implications of increasing autonomy

During this session, the speakers addressed the legal and ethical implications of increasing autonomy in weapon systems, with a focus on compliance with international humanitarian law (IHL) and questions of accountability.

Using the ICRC's working definition of an autonomous weapon system, the first speaker reiterated that any such weapon must comply with IHL rules on the conduct of hostilities, suggesting that compliance might differ depending on the specific weapon and its context of use. Among the key challenges to IHL compliance, the speaker stressed that it was questionable whether a weapon system could be programmed to distinguish between civilians and combatants, and in particular whether the definition of a civilian could be converted into computer code. Likewise the speaker questioned the ability of a machine to apply the rule of proportionality in attack, which involves a balance of different values and appears to require uniquely human judgement.

The speaker suggested that national legal reviews were important to ensure compliance with IHL, but also expressed the concern that overemphasizing domestic legal reviews could provide a legal pretext for weapons that should not be developed in the first place. The speaker stressed that an international instrument prohibiting or limiting those weapons would be desirable, especially in light of other potential risks, such as lowering the threshold for the use of force. The speaker suggested, however, that greater distinction was needed among the types of systems that would raise challenges for IHL compliance, and remarked that much of the current discussion was based on the assumption that "fully autonomous weapon systems" might be possible in the future, which made it difficult to draw definitive conclusions.

In conclusion, the speaker questioned whether IHL should be the only criterion to consider when judging a new weapon system. In that respect, the speaker highlighted a number of questions for further discussion at the international level, including: the need to develop a precise definition of autonomous weapon systems as a precondition for discussions concerning their legality and eventual prohibition; and the need to encourage more developing countries to join debates about autonomous weapon systems, with a view to developing a widely accepted international instrument to regulate those weapons.

The next speaker noted that IHL does not contain a general prohibition of autonomous weapon systems and that, given the wide range of potential types of those weapons, an assessment of their legality cannot be made in the abstract. The speaker also stressed that IHL rules on the conduct of hostilities are addressed to the parties to the conflict, more specifically to human beings. While the primary subjects of IHL are States, the IHL rules of distinction, proportionality and precautions in attack are addressed (implicitly or explicitly) to the individuals who plan and decide upon an attack. Those rules create obligations for human combatants and fighters, who are responsible for respecting them and would be held accountable for violations.

The speaker went on to describe three different stages where human control could be exercised in relation to autonomous weapon systems, i.e. in the development, deployment and operational phases. A key question was raised as to whether human control in the first two stages would be sufficient to overcome minimal or no human control at the last stage, where the weapon system autonomously selects and attacks targets. As had been discussed in previous sessions, the speaker emphasized that many defensive systems were already capable, after initial activation by a human operator, of autonomously selecting and attacking targets (third stage) to defend ships, vehicles or ground bases against incoming missiles or rockets.

The speaker also discussed the challenges posed by autonomous systems to legal reviews of new weapons, including the absence of standard methods and protocols for testing and evaluation to assess the performance of those weapons, and the possible risks associated with their use. Questions were raised regarding: How was the reliability (e.g. the risk of malfunction or vulnerability to cyber attack) and predictability of the weapon tested? What level of reliability and predictability were considered necessary?

On the question of a possible accountability gap with autonomous weapon systems, and considering only so called "fully autonomous weapon systems" (with no human oversight), the third speaker began by examining criminal liability, asserting that the subjective mental element (*mens rea*) – which required proving the intent of a human programmer or operator – could be hard to fulfil in some situations. Using the example of a direct attack on civilians by an autonomous weapon system, the speaker explained that, applying the International Criminal Court's *mens rea* standard, one would need to prove that the programmer or operator of the weapon intended it to directly attack civilians or knew with certainty that such a violation would occur. Applying the Additional Protocol I and customary-criminal-law standard of "wilful killing" of civilians, it would be sufficient to prove that the programmer or operator wilfully accepted the risk that the machine might take the wrong targeting decision and directly attack civilians. The speaker recalled that the standard was one of *indirect* intent (*dolus eventualis*), which all States party to Additional Protocol I were bound to apply, and in that respect, the so-called accountability gap seemed less wide.

The speaker then turned to the law of State responsibility, which, it was argued, is not challenged by the development of autonomous weapon systems since, unlike criminal law, it does not require a subjective element. The speaker said it would be sufficient for the act to be *objectively* attributable to the State and that attacks carried out by autonomous weapon systems would not pose any specific problems with regard to attribution in that respect. If faithfully implemented, the framework of State responsibility could have a significant deterrent effect, the speaker added, since it forced States to provide guarantees of non-repetition and full reparation, including compensation for victims.

During the discussion, there was a debate about the role of legal reviews of new weapons (as required by Article 36 of Additional Protocol I) in addressing issues raised by autonomous weapon systems. One participant stressed their importance in ensuring the compliance of any new weapon with IHL but noted that few States currently carried out such reviews. Another participant pointed out that the process allowed for very limited transparency, owing to the sensitive nature of the information, and that it would be difficult to imagine the sharing of review results among States. Finally, another participant argued that, while important, legal reviews did not provide a solution to all the questions raised by autonomous weapon systems, including the implications for international security and stability.

One participant raised the question of whether autonomous weapon systems might be considered indiscriminate weapons. A speaker responded that the question would likely depend on the specific weapon system and the context of its use. For example, the speaker noted that existing autonomous weapon systems, such as rocket- and missile-defence systems, were used to perform a single task in a specific, contained and "uncluttered" environment where there was little or no risk of encountering protected objects. However, one might imagine an autonomous weapon system designed to be deployed in a complex, "cluttered" environment, i.e. where it was likely to encounter civilians and civilian objects, yet was incapable of distinguishing military objectives from civilians and civilian objects; in such a case, the autonomous weapon system would be considered an indiscriminate weapon. One participant emphasized that military commanders were not calling for increased acquisition of autonomous weapon systems, because that would go against their aim to ensure control over the battlespace.

There was also discussion about the notion of "attack" under IHL. A participant emphasized that there was no distinction, from a legal perspective, between a "defensive" and an "offensive" weapon system, since both were used to carry out attacks. The participant raised the question of what constituted an attack, and at what stage an assessment of the legality of the attack must be made, i.e. to ensure the attack is discriminate and proportionate. In other words, would the assessment be made at the point of activation of the machine, or prior to each individual attack? A speaker responded that each use of force must be in compliance with IHL, but that for pre-planned attacks, the legal assessments were made at the planning stage through tools such as collateral-damage estimates, also taking into consideration the available means.[7]

5. Human control

This session focused on human control over weapon systems and the use of force, thus providing an alternative approach to analysing autonomous weapon systems from a purely technical perspective.

The first speaker explained the concept of meaningful human control over individual attacks, arguing that such control was a requirement for IHL compliance, as well as a useful means of determining the boundaries beyond which autonomous weapon systems would be unacceptable (i.e. without meaningful human control). The speaker highlighted the key elements of meaningful human control as follows:

- information on the military objective;
- understanding of the technology, including predictability and reliability;
- information on the context, including time and space limitations;
- analysis and understanding of how the technology and the context would interact, including risks to civilians;
- human judgement and the potential for timely action; and
- a framework of accountability.

The speaker emphasized that the rules of IHL applying to attacks were addressed to human beings ("those who plan and decide upon an attack"), and therefore the obligation to apply the rules rested with humans. Machines could not apply the law, but must carry out operations in line with legal judgements made by humans. The speaker raised concerns that increasingly autonomous weapon systems risked expanding the notion of attack, which in the view of the speaker was a unit of military action limited in time and space, and over which individual human legal judgements were required by IHL. The speaker said that existing weapon systems were mostly constrained in their functioning in time and space, but that relaxing those temporal and spatial limits would necessarily decrease human control over attacks, as would allowing machines the latitude to set their own objectives. For example, an autonomous weapon system that "hunted" for targets over a wide area would raise concerns about human control over attacks, owing to the lack of knowledge about where and when each attack would occur.

The second speaker offered another concept of human control based on the decision-making cycle that surrounds an attack. Using the targeting process of the NATO Joint Targeting Cycle as an example, the speaker explained where human control could intervene in that process, and related the level of human involvement in this process to the level of autonomy in a particular weapon system.

[7] Note: It remained unclear from the discussion whether the moment of activation of an autonomous weapon system would constitute an attack, or only the moment when the system used force against a target. This would have an impact on when the commander must carry out an assessment of proportionality and determine which precautions to take, and the related question of whether it would be possible to effectively take such measures at the point of activation of the weapon system.

The speaker explained the various stages of pre-planning and assessment that take place in the targeting cycle before and after the use of force. The speaker emphasized that, for existing autonomous weapon systems which select and attack targets without human intervention, human control was exerted in the phases of the targeting process that preceded the weapon's activation, and during which decisions were taken to select and develop targets, and to select a specific weapon for a particular task in a certain context, among others. The speaker said that human control was also exerted through operational constraints, such as limitations in time and space, which were placed on the use of the weapon before the moment at which the system selected and attacked targets autonomously.

The speaker asserted that, for existing autonomous weapon systems, although there might be no direct human control over the system's critical functions of selecting and attacking targets, the targeting process as a whole was largely human-dominated. However, the speaker cautioned that, with rapid technological advances, there might be a boundary beyond which machines were given too much control over the targeting process, and human control would then be overridden. For example, the speaker said, weapon systems that adapted or learned, developed their own objectives and target lists, and changed their functioning could present such a risk.

The third speaker introduced ethical and moral considerations related to increased autonomy, with a focus on the potential risks and unintended consequences posed by autonomous weapon systems. The range of risks mentioned, resulting from those weapon systems not functioning as intended, included: fratricide, civilian harm, unintended initiation/escalation of conflict, hacking, spoofing and "normal accidents". The speaker emphasized that the magnitude of those different risks would be significantly affected by the characteristics of the specific weapon and its context of use, including: the time of operation and geographical range of the weapon; the potential damage (related to the munitions the weapon fired); the size of the magazine (i.e. the quantity of ammunition); the ability of, and time taken for, a human operator to shut down the system; the number of weapon systems deployed; and the number of "contacts" with potential targets.

The speaker explained that failures of autonomous weapon systems would certainly occur, as with any complex system, and that high-frequency use would still lead to a significant number of failures, even with measures taken to mitigate the risk of failure. The Patriot missile-defence system was cited as an example of the failure rate in autonomous weapon systems; out of 13 engagements in a particular operational period, the system apparently had resulted in two fratricide incidents. The speaker added that autonomous systems were intrinsically unpredictable in their operation, and that such unpredictability would be exacerbated further where such systems came into contact with other autonomous weapon systems.

In order to minimize unintended risks, the speaker argued, it was essential to: retain human control over critical operations of weapon systems; ensure that human moral agency was retained in targeting decisions; and ensure that systems with some degree of autonomy were designed with a fail-safe procedure (i.e. deactivation) as a last resort.

During the discussion, several participants stressed that human control over any weapon system was not only essential from an ethical and legal point of view, but also from a military operational perspective.

One participant asked whether increasing autonomy led to a decrease in (meaningful) human control. The speakers responded that this was not necessarily the case, but that it would depend on the specific function of the weapon system and the context in which it was being used. One participant expressed a preference for the term "human control" rather than

"meaningful human control," since, in their view, human control over a weapon system was either present or it was not.

Another participant asked whether existing autonomous weapon systems operated with meaningful human control. Speakers responded that certain constraints enabled human control to be exerted, in particular, time-and-space restrictions, human selection of the specific target, and knowledge of the environment within which the weapon system operated. By comparison, one speaker pointed out that concerns about human control would be raised in situations where the specific location in which force would be used was not known to the user of the weapon system. Another participant added that the distinction between a legitimate target (i.e. military objectives) and protected objects (i.e. civilian objects) could vary over time and depending on the context. Therefore, in order to maintain human control and compliance with IHL, it was essential to control the space and time over which weapon systems operated.

6. Addressing the challenges raised by increasing autonomy

The final session of the meeting discussed potential approaches to the challenges raised by increasing autonomy in weapon systems, and considered how to ensure that human control over the use of force is maintained.

The first speaker addressed the issue from a military decision-making perspective, arguing for the need to develop an evolving partnership between humans and machines. The speaker distinguished between automated and autonomous weapon systems, arguing that the former were programmed to a pre-defined set of rules with a predictable outcome, while the latter would be capable of deciding on a course of action from among a number of alternatives. The speaker added that the overall operation of autonomous weapon systems would be predictable, but that individual actions might not be. Based on that distinction, the speaker said that it was doubtful whether such an autonomous weapon system could ever replace the need for decision-making by a military commander.

The speaker explained that military decision-making always had an intuitive component as well as an analytical one, and that it was guided by: professional judgement gained from experience, knowledge, education, intelligence and intuition. It would be difficult, therefore, to envisage the intuitive part of decision-making being carried out by a machine. The speaker stressed that it would always be necessary to have a human-led process for high-stakes decisions, such as targeting.

Nevertheless, the speaker cautioned against a pre-emptive prohibition of autonomous weapon systems, saying that it would hamper ongoing research on growing autonomy in weapon systems with the aim of increasing precision and target discrimination, and for defensive purposes. The speaker added that States should focus on their current obligation to put in place a robust legal review process to ensure that new weapons complied with IHL.

The second speaker provided a perspective on the development of autonomous weapon systems in Russia, explaining that the Russian Ministry of Defence used the term "combat robot" to describe a "multifunctional device with anthropomorphic (humanlike) behaviour that partially or fully performs functions of a human during particular combat missions". The speaker explained that Russia had recently been investing more and more in the development of robotics, including autonomous systems, in both the civilian and military spheres, and that, in September 2015, the Russian Defence Ministry had developed a Comprehensive Policy "Programme for Development of Advanced Military Robotics up to 2025 with Forecasts until 2030," reflecting the main trends in the development of robotic systems for military purposes. The speaker explained that all existing Russian Army systems were remote-controlled. However, some of these could be used in a partially autonomous

mode and, in the future, those systems could be reprogrammed to operate with an even higher degree of autonomy.

The speaker described a number of existing robotic systems, noting that Russia's fleet included unmanned aerial vehicles, such as the Orlan-10 (an unarmed reconnaissance aircraft) and the Eleron-3SV (a reconnaissance and electronic jamming aircraft), as well as the unmanned ground vehicles Raznoboy and Berloga-P, which are used for remote controlled radiation and chemical monitoring. In addition, the speaker described future models, such as the Cobra-1600 Light Sapper Robot (for remote-controlled reconnaissance and bomb disposal), to be deployed in 2016, and several systems being tested, such as the Uran-6 minesweeping system, and the Uran-9 unmanned combat ground vehicle, which will be designed for combined combat and reconnaissance operations as well as fire support. The speaker explained that some systems were at the testing stage, such as the Platforma-M, which is designed to carry out rescue missions and could also be used to lay smoke screens and plant mines. The speaker added that these systems could be used to replace personnel and to protect borders. However, the speaker emphasized the importance of compliance with IHL, an issue taken seriously by Russia.

The speaker also highlighted some risks posed by autonomous weapon systems, including the potential for accidental attacks due to loss of communication, jamming, interception, or cyber-security failures. Most notably, however, the speaker stressed that the development of autonomous weapon systems might lead to a new arms race and substantially increase the risk of armed conflict.

The third speaker provided a different perspective, with five proposals for framing discussions on autonomous weapon systems at future meetings within the framework of the CCW, namely to:

- avoid trying to define "autonomous weapon systems" and rather think about autonomy *in* weapon systems, with a focus on critical functions.
- draw lessons from existing weapon systems with a high degree of automation in their critical functions. Understanding the parameters and boundaries that are not problematic from a legal and ethical perspective would help to identify developments that might raise concerns.
- increase attention to the implications of both machine-learning systems – in particular, the implications for unpredictability – and cyber weapons, since the effects of autonomous weapon systems might not be limited to kinetic effects.
- consider the implications of alternative development pathways for autonomous weapon systems, in particular the use of "off-the-shelf" technology to enable the weaponization of increasingly autonomous civilian robotics technology by individuals or non-State armed groups.
- reframe the CCW discussions so that the issue centres on the role of the human rather than the technology itself. The concept of human control provides a common language for States to determine the degree and type of control and oversight over weapons and the use of force that is required.

There was debate among the participants on the need for specific regulation or prohibition of autonomous weapon systems. Broadly speaking, there were three approaches proposed by different participants, which could be pursued individually or in parallel.

Firstly, one participant argued that the existing IHL framework was sufficient to address the relevant issues, and that States should focus on better implementation of legal reviews of new weapons (as required by Article 36 of Additional Protocol I). Another participant said that a disadvantage of such an approach was that the legal assessment of autonomous weapon

systems was open to different interpretations, that there was a lack of transparency concerning legal reviews, that such reviews were unilateral, and that the said approach could present a risk of legitimizing autonomous weapon systems.

Secondly, as a participant explained, another approach would be to develop a new instrument within the framework of the CCW to regulate or prohibit autonomous weapon systems. A key aspect of such a process, in the view of the participant, would be agreement on a definition of autonomous weapon systems that were to be regulated or prohibited. The participant added that such an approach could be pursued in parallel with increased attention to national legal reviews.

Thirdly, another participant proposed an IHL-compliance-based approach to the issue, which would build on existing obligations in order to better understand where developments in autonomous weapon systems might raise concerns. The participant said that there was some consensus on the need for human control, or the involvement of humans, in weapon systems and decisions to use force, but that there was currently a need to determine the kind and degree of control that was necessary to comply with existing IHL. That analysis would help to draw a line between autonomous weapon systems that might be acceptable, including some existing systems, and those that might need regulation or prohibition.

Another issue raised during the discussion was the lack of clarity about whether there was a genuine distinction between “highly automated” and “autonomous” weapon systems. One participant said that it would be possible to define autonomous weapon systems of concern and “draw a red line” for those weapons that must be prohibited. However, another participant noted that highly automated weapon systems raised similar legal and ethical questions, and that “fully autonomous” weapon systems might never exist.

Another participant cautioned against focusing solely on definitions, calling for a more proactive approach to addressing the challenges, and pointing out that the debate on definitions had already been going on for many years, while the CCW process had lagged behind rapid technical developments in the field. One speaker responded that there was a need to delineate the scope of the discussion, but that lessons could be drawn from case studies of autonomy in existing weapon systems. Another speaker added that a focus on human control would enable a better understanding of the requirements under IHL and a means to develop more concrete proposals to address weapon systems of concern.

PART II: SELECTED PRESENTATIONS

SESSION 1: CHARACTERISTICS OF AUTONOMOUS WEAPON SYSTEMS

Characteristics of autonomous weapon systems

Speaker's summary

Dr Martin Hagström, Swedish Defence Research Agency, Sweden

The subject of autonomous weapon systems has drawn increasing attention in recent years. Although the debate about such systems has grown significantly since the publication of the US policy document entitled *The Role of Autonomy in DoD* [Department of Defense] *Systems*, autonomous weapons have been around for more than a century. During the First World War, aerial torpedoes were developed. These were ground-to-ground guided missiles which, after launch, were completely autonomous. During the Second World War, the development of guided missiles continued, and today weapons with a high degree of automation, or self-guidance, can be found in the inventory of most States.

There are several reasons for the ongoing debate about autonomous weapons. One concerns the word “autonomous”, which implies self-governance and decision-making. Weapons are used in armed conflicts, and the use of weapons leads to people’s death. Therefore, the question arises: Will autonomous weapons make decisions over life and death? However, the anthropomorphic use of words causes confusion. Machines, as we know them today, and in the foreseeable future, will remain machines. The “autonomy” of autonomous systems is created by complex computer programs. Computers compute, and the results, however amazing, are the result of calculations. Human attributes, in contrast, are different from machine characteristics; many of the words used to describe characteristics, such as “learning”, “autonomy” and “decisions”, have a completely different meaning when referring to machines as opposed to humans.

In technical contexts, the word “autonomous” is used to describe a system which, without direct influence from an operator, can act in an unknown environment or handle unexpected events. Engineers use the word “unexpected” to describe events in the environment that are not foreseen in detail: for example, exactly how a road turns, and how the wind speed varies over time. That the road can turn and the wind speed vary is, however, anticipated and described comprehensively in a model of the environment. Aircraft autopilots, for example, are designed to handle gusts and changes in the load and centre of gravity, the details of which are unknown but which are, in the model, expected variations in the environment. These changes are new conditions which are in some sense anticipated. What distinguishes an automated system from an autonomous system is merely the perception of the complexity of the functions that are automatic. The word “automatic” is often used for individual functions, but “autonomous” is used for an assembly of several “automatic” functions. There is no clear boundary between what is perceived as an “automatic” function and an “autonomous” system. A well-known and familiar technology is more often referred to as “automatic”, while new automated technology is labelled “autonomous”.

The automation of a function requires knowledge and understanding of the task to be performed. The piloting of an aircraft today is considered simple automation. An autopilot,

just like a human pilot, needs to compensate for small unforeseen changes in conditions, such as wind gusts. The autopilot must "understand" the aircraft's behaviour, e.g. how the aircraft reacts when a rudder is turned. This "understanding" is a description of the world the aircraft operates in. It can be a mathematical description, or model, of the relations between actions and reactions. The aircraft will go up if the elevator (the rudder controlling elevation) is turned up, and left or right if the sideways rudder is turned left or right. The model also describes the aircraft's dependency on gravity, wind, Earth rotation, etc.

The model describes the universe in which the system acts – its design space. Every autonomous system is designed to act within that space. There is always a model defining the system's universe. It can be explicit, with mathematical descriptions of known physical laws, as in the aircraft control example, or implicit, such as a black box. The black box can be the result of a complex process where mathematical methods have been used to design a model without explicit human understanding of all details. This is the typical result of "machine learning", another anthropomorphic use of words. Machine learning, along with the recent term "deep learning", is a method of identifying patterns and structures and storing them in a model. The model can be the basis for an autonomous system, which can then act within the model's universe, the system's design space. Once a system is placed outside its design space (i.e. owing to a truly unforeseen event), its response is unpredictable by definition. From a human perspective, the response might be good, or it might be "bad", but it cannot be foreseen. How to make the design space as big as possible, i.e. in some sense to foresee as much as possible, is an engineering challenge.

There are several reasons for introducing a higher degree of automation in military systems. They include increasing performance and reducing costs, in addition to reducing operator risks. Superior performance in terms of speed is a key factor in an armed duel. Humans have a limited ability to respond to rapid sequences of events, and when information is collected and processed, and decisions must be taken within fractions of seconds, machines are usually better suited to the task than humans. There is a contradiction between the requirement for human control of a weapon's effect and the weapon's performance. In military contexts this is not a new situation. In military operations, there is often limited time for decisions, and situations cannot always be thoroughly analysed during combat. Therefore, the decision on the use of force is a well-defined process based on doctrines, methods of warfare and rules of engagement. The use of a weapon must be preceded by analysis and the development of doctrines, manuals and training programmes, and the more complex the system, the more extensive will be the preparations that are needed.

One source of the arguments against autonomous weapons is their perceived unpredictability, since weapons that can kill should not be unpredictable. The focus on unpredictability is due to the complexity of autonomous systems. However, complex systems are used in other areas where failures might have catastrophic consequences. Systems which, if malfunctioning, can cause danger, harm or even death are called "safety-critical systems". Typically these are aerospace, nuclear-power, rail and weapon systems. The performance, use and development of these systems, which can threaten human safety, are regulated in many respects by legislation or common standards. The failure, or unintended effects, of a complex technical system is seldom the result of a single cause. Since a complex system is developed over a long period by many people, manufactured by others and often operated by an organization different from the producer, there are many possible reasons for an undesired effect. In the case of failure, it can be difficult to trace it back to a single reason or person responsible. Therefore, legislation and standards for the development and use of such systems exist in many areas. Safety-critical systems need a thorough analysis of their technology, intended use and possible unintended use.

It is difficult to imagine a definition of criteria or characteristics that would aim to draw a line between an acceptable level of automation and an unacceptable level of autonomy from a

technical perspective. Since the “level” of autonomy is not well defined, it will be dependent on the situation and the system. Requirements might be formulated with a focus on system reliability, procedures for development, and the development of doctrines and training programmes. Such requirements do not depend on a specific technology, but on performance, controllability and use.

All we fear, and all we hope for, has already been written about with respect to autonomous weapons. The debate is partially influenced by psychological driving forces that also have close links to the use of anthropomorphic concepts. Fiction, in particular science fiction, has described many of these driving forces in a long line of books. The fear that robots could be callous and merciless killers appears repeatedly in discussions. Science fiction provides a guide to understanding the elements of the debate, and an overview of the threats conjured up by those who are opposed to autonomous weapon systems, along with the opportunities they present. The *Terminator* movies, with the artificial intelligence, Skynet, are of course the typical example of the artificial intelligence threat, but as early as 1953, Philip K. Dick wrote a novel entitled *Second Variety* (which became the film *Screamers*), about fully autonomous weapons that are self-learning and that ultimately threaten all of humanity. The movie *Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb* (1964), which was based on the novel *Red Alert* (1958), deals with the problems of the arms race and the possibility of an accidental nuclear war caused by the automation of weapons. The list goes on. Every conceivable threat and opportunity has been described in the literature.

Focusing the debate on autonomous weapon systems:
A new approach to linking technology and IHL

Speaker's summary

Lt Col. Alan Schuller, Stockton Center for the Study of International Law,
US Naval War College, USA[1]

The Stockton Center for the Study of International Law has embarked on a year-long project to link international humanitarian law (IHL) to the technology and military application of autonomous weapon systems. Our goal is to create an objective report that can be used by researchers as well as by policymakers and practitioners. Today I would like to share with you some thoughts regarding a different approach to evaluating the characteristics of autonomous weapon systems. I invite you to challenge your assumptions regarding the development of technology and how the law will apply to these systems.

With regard to defining autonomy, we must stop trying to describe the category of autonomous systems as a whole and focus instead on delineating what combinations of autonomy would potentially be unlawful. Simply put, "autonomous weapon systems" is an overly broad category when attempting to devise all-encompassing legal principles. The technology is too diverse to describe succinctly yet comprehensively from a legal perspective. Further, "select and engage" may be useful in describing a segment of automation that we should look at carefully because of its operational significance, but it is less helpful in defining a category of automation that is legally objectionable. Instead of attempting to describe and regulate the entire possible spectrum of autonomy, therefore, we should establish best practices, delineating distinct combinations of autonomous technologies that cause us particular concern.

A simple construct to frame the discussion is the "OODA loop" (a decision cycle of observe, orient, decide and act). I am not referring, however, to where a human might be placed vis-à-vis the cycle, but instead to which puzzle-shaped pieces from the loop have been delegated to computers. For it is the ever-increasing surrender of portions of the OODA loop to machines which may ultimately lead to issues with IHL compliance. In this context, pieces might consist of *authority* (e.g. in the precise programming or learning capacity of the computer) and/or *physical* capabilities (e.g. the ability to loiter for a long duration). As such, the critical issue bearing on IHL compliance may not be whether the machine "selects and engages" without human intervention, but rather whether it has been granted some critical combination of functions that effectively delegate the *decision to kill* from human to machine. For if a machine is able to precisely identify (both in terms of the nature of the object and its location in time and space) and attack a narrowly defined target provided to it by a human, the machine did not *select* the target as the object of the attack; the human did. As such, the question necessarily becomes: can the actions of the machine reasonably be traced back to the decision by a human to attack the target or class of target?

The decision to kill, which invokes analysis under IHL, is without question a human's burden. This decision inherently implies IHL analysis deriving from the potential use of force. This is not a digression into philosophical inquiry; it is instead in this case a technological evaluation. Have we ceded so much autonomy – so many pieces of the OODA loop, or more importantly, just the *right* combinations – that we can no longer say that a human functionally

[1] Military professor, Stockton Center for the Study of International Law, US Naval War College, and fellow, Georgetown Center on National Security and the Law. The views set forth herein were expressed in my personal capacity and should not be attributed to any of my institutional affiliates, including the US Department of Defense and the US Naval War College.

decided to kill? Importantly, this does not mean that a human was temporally proximate to the moment of kinetic action. We could avoid functionally delegating the decision to kill, for example, by means of carefully tailored computer programming or a control tether.

So how do we prevent ourselves from functionally delegating that which we may not delegate? **Predictability is the key.** But not all aspects of the system must be predictable. There is, of course, great potential military advantage to be gained by providing advanced machine learning, for example, to those aspects of a machine which either do not bear on IHL compliance or do not combine with other autonomous features to functionally delegate the decision to kill. Those aspects of the system, however, which in combination may affect our ability to reasonably predict compliance with IHL, are where we must focus our evaluation. Like most IHL requirements, our ability to predict the machine's actions must be based on a standard of reasonableness. A lower standard would encourage us to unlawfully relieve ourselves of the obligation to comply with IHL by blaming computers for violations. A higher standard would be unreasonable, given the complexity of computer programming as magnified by the fog of the battlefield.

Predictability cannot diminish past the point where we can reasonably say that a human was in control of compliance with IHL. Importantly, this is *not* the same standard as physical human control over the actions of the machine itself at the time of lethal kinetic action. It also does *not* mean that a human made a decision on IHL compliance that was temporally proximate to a lethal attack. It means that we can reasonably predict what decision the system will make and that we are reasonably certain the system will comply with IHL. If we can reasonably predict compliance, we can maintain effective control despite our level and type of interaction with the machine at the time of lethal action. If, on the other hand, we cannot reasonably predict whether the machine will comply with IHL, it is potentially *unlawfully autonomous*.

We must stop trying to draw a line between autonomous and automated. This is a futile effort that attempts to paint over infinite shades of grey with a facade of order. It is also likely a quest to know the unknowable. Most importantly, there is no legal tipping point inherent in these descriptions because they are non-linear at best and arbitrary at worst. More automation does not always lead to autonomy or to legal objections, and broad-brush categorizations are therefore not useful in describing specific combinations of autonomy which are legally problematic. Instead, we must focus on whether specific combinations of pieces from the OODA loop have been surrendered to a computer such that we have functionally delegated the decision to kill to a machine, since a human can no longer reasonably predict compliance with IHL.

ICRC working definition of autonomous weapon systems

Speaker's summary

Dr Neil Davison, Scientific and Policy Adviser, Arms Unit, Legal Division, ICRC

Note: For a summary of the issues raised in this presentation, see Section 2 of the ICRC's background paper in Part III of this report.

SESSION 2: AUTONOMY N EXISTING WEAPONS

Missile defence systems that use computers: An overview of the Counter-Rocket, Artillery and Mortar (C-RAM) system

Speaker's summary

Dr Brian Hall, Joint Chiefs of Staff, Department of Defense, USA

The speaker presented an overview of the technical operation and military utility of a general-category semi-autonomous weapon system, specifically, the Counter-Rocket, Artillery and Mortar (C-RAM) system. The presentation covered why and how this system was developed, how it functions, and whether the system has performed as intended.

The speaker emphasized that functions and features of the C-RAM system were developed specifically in relation to its military role. He said that the content of the presentation should not be misconstrued as reflecting broader principles related to functions and features to be applied to other weapon systems, including systems that were computer-aided and -enabled. For example, just because a particular function might be important to C-RAM's operation, that did not mean that the same function would be important for all weapon systems displaying an element of autonomy. What functions were important for a particular system depended upon that system's purpose – in the case at hand, the protection of military forces, civilians and infrastructure.

The speaker showed a short video to demonstrate that C-RAM was actually a system of systems. That helped the audience to better understand that its capability encompassed not just the Land-based Phalanx Weapons System (LPWS), but an integration of various threat-detection, threat-warning, command-and-control and engagement features. That design configuration was the direct result of the original operational need identified in 2004 during multinational operations in Iraq. The need had been translated into a capability designed to react quickly and effectively with greater precision and accuracy than any existing methods to counter the rocket, artillery and mortar threat to soldiers and civilians.

The speaker then explained that the C-RAM technology was not new, but had been used by the US Navy since the early 1960s as a terminal defence against anti-ship missiles. Further development of the land-based version of C-RAM complied with the US Department of Defense acquisition and procurement processes. In those processes, defence acquisition professionals fully understood the need for a new system and conveyed that requirement to industry. Emphasis was placed on the many types of professionals ensuring that any new weapon system met valid operational requirements, worked as intended, could be designed and used in safety and complied with legal convention. To dismiss any notion that the US acquisition and procurement processes were simple, the speaker showed the audience the current graphic describing the complex US Integrated Defense Acquisition, Technology, and Logistics Life Cycle Management System. The speaker emphasized that, embedded in any system's life cycle were numerous, recurring weapon review boards and weapon system safety reviews demonstrating legal compliance and adherence to safety standards.

The presentation then included discussion emphasizing C-RAM as a mix of human decision-making and automation encompassed within a system-of-systems architecture and a concept of operations. Both of those clearly showed C-RAM to be a semi-autonomous weapon system with inherent safeguards to prevent unintended use.

In closing, the speaker noted that the integrated system had protected people and property by shooting down missiles and mortars in hundreds of attacks since 2005. It did that by leveraging the advantages derived from the use of computers and human abilities. Specifically, automation had been used to optimize the timing and increase the precision of fires used for tasks within the overall protection mission. C-RAM had simply worked as intended.

Missile- and rocket-defence weapon systems

Speaker's summary

Gp Capt. Ajey Lele (Ret'd), Institute for Defence Studies and Analyses, India

Various defensive mechanisms are available to guard against incoming missile attacks. This presentation discusses several important missile-defence systems and their efficacy for twenty-first-century warfare. The presentation highlights the autonomous nature of such systems and the debates regarding their future development.

Broadly, autonomous weapon systems are "fire-and-forget" systems which, once activated, select and engage targets on their own without any human intervention. A given weapon system may be either offensive or defensive; however, owing to the nature of warfare, fully autonomous systems are expected to belong to the defence category. It is not possible for a missile system to choose a target on its own, because no machine can decide why, when, where and how to start a conflict unless, and until, it is programmed to do so. Technically and technologically, if any missile is in attack mode without exact knowledge of the target activated, then its seeker is likely to search for the target in its field of view and would eventually get confused. In this process, it would run out of fuel and make the self-tasked mission unproductive.

Presently, the known and successful defensive systems, and those under development, that are fully autonomous in selecting and attacking targets are: counter-rocket, artillery and mortar systems, such as Iron Dome, and anti-missile systems, such as Terminal High Altitude Aerial Defense (THAAD), S-400, etc.

For any system, once the target has been identified, the rest of the work is done mostly by the guidance system. This combines navigation-satellite and path-computing units with a guidance control system. The navigation-satellite and computational unit calculates the path and trajectories, and the guidance system then controls the operation of the interceptor missile. The incoming threat gets detected by land-based radar for short-range targets, and by radar satellite for threats coming from a distance. The radar sends data to the control unit, which calculates the threat trajectory, and, on that basis, sends the signal to the most appropriate unit that will effectively intercept the incoming threat. The "artificially intelligent" controller controls this whole process. The controller, guidance and seeking systems are able to differentiate between friendly aircraft and an incoming threat.

It is important to note that autonomy cannot be absolute; there may be either a low or a high level of autonomy. Interception can be either endo-atmospheric or exo-atmospheric – that is, it can take place either inside or outside the Earth's atmosphere. Anti-missile defence systems would be kept in a "ready state" depending on the threat perception. It is possible that in some cases they would always remain in operational mode. The nature and performance of the defence system also depends on what type of threat the systems have been designed to address. The performance of various autonomous missile defence systems is better if they are designed to address an incoming ballistic-missile threat. The existing level of technology shows limitations in addressing cruise-missile threats. Also, in case of a saturation raid, the effectiveness of such systems, even against ballistic missiles, becomes degraded. The ongoing technological developments in cyber weapons and other non-kinetic weapon fields could emerge as better options for addressing incoming missile threats in the future. Also, the limitations of missile-defence systems in countering directed-energy weapons, such as lasers, are becoming more evident.

A good example of a short-range system is Iron Dome, which is a counter-rocket, artillery, and mortar system capable of intercepting multiple targets from any direction. The autonomous guidance and control system of the Iron Dome is capable of intercepting only those targets which represent a high-priority threat according to the system configuration. In addition, this system is able to successfully intercept 90% of incoming threats from a 4-km range. For threats coming from a longer distance, the most suitable missile-defence system currently is THAAD. When a threat missile gets launched, an infrared satellite detects its heat signature and sends an early warning and other useful real-time tracking data to the ground-based system through a communications satellite. When the threat is confirmed by analysis (with no human involvement), the appropriate command gets delivered to sensors and weapon systems. After that, the long-range radar detects and tracks the missile for some time to improve accuracy. The tracking data helps to calculate the near-accurate trajectory of the incoming threat missile. Among the group of batteries available to address the threat, the most effective interceptor battery is engaged and carries out the interception. The complete process of killing the missile is fully autonomous in nature and hypothetically has very high efficiency.

The performance of various autonomous missile-defence systems can be constrained owing to a number of factors. These include the technical configuration of the computational units, seeking radars, control algorithms and missile controls, the speed of communication between different units and how tracking the target affects system performance.

Apart from missile-defence systems, there are some other autonomous weapon systems involving rocket technologies. These are space-based autonomous systems which could be used to target space-based systems, as well as targets on Earth. It is important to note that such systems at present are mostly in the realm of theoretical possibility; however, it is possible that States could make such systems operational in the near future.

Currently, the trend in the development of missile-defence and space-based systems is toward increasing autonomy. Technically, 100% autonomy could be considered a myth; however, the degree of system autonomy is expected to increase many times over in the near future. The ability to effectively control missile-defence and space-based weapon systems would depend on a number of factors. Missile-defence capability is emerging as a cornerstone of strategic doctrine for some States. Also, there are situations where missile-defence systems are used more for geopolitical reasons, and such systems are also known to have “deterrence” potential. Unfortunately, all of the nine nuclear-weapon States in the world are known to be increasing their nuclear arsenals at present. Similarly, investments in missile-defence systems and space-based weapon systems are also expected to rise. All of this would demand increasing autonomy in such systems.

Sensor-fused munitions, missiles and loitering munitions

Speaker's summary

Dr Heather Roff, Senior Research Fellow, Department of Politics and International Relations, University of Oxford, UK, and Research Scientist, Global Security Initiative, Arizona State University, USA

The presentation sought to answer four main questions: (1) What is the state of military weapon technology today? (2) Where do we see autonomy in "critical functions"? (3) What is the trajectory of autonomy in weapon systems? (4) Where will we likely see autonomous weapons develop?

1. The state of military weapon technology today

In assessing the present state of military weapon systems, I looked at the top five weapon-exporting countries (the USA, Russia, China, Germany and France) and surveyed their presently deployed missile and bomb arsenals. These five countries make up 74% of the world's arms trade, and as such are leaders in weapon development and export. The data consist of over 230 weapon systems.

The data suggest that most advancements relate to homing, navigation, target acquisition, target identification, target prioritization, 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 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 an attack.

2. 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 or direction by a human operator, or something more robust, such as cognitive capacities in making a "decision". For the sake of this presentation, almost all data was coded as binary (as either a zero or a one), so as to move away from whether the system was "autonomous" or "automatic". 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 one. What is more, close-in defensive weapon systems are also capable of sensing a target, prioritizing that target and firing on it without the intervention of a human operator; these are also coded as a one.

That said, in current weapon systems, the "selection" of targets may be better thought of as "detection". Present-day systems have various sensor capabilities that allow them to perceive their surroundings and then to recognize 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 limited cases, such as anti-ship missiles, these

systems are also utilizing various sensor capabilities to navigate, locate and identify targets (ships). Once there, they are able to select from among various identified targets, but it appears that they do so by prioritizing, ostensibly given some sort of predefined criteria. Loitering munitions may or may not have a "human in the loop" to select a target.

3. Trajectory of autonomy in weapon systems

The trajectories of autonomy in weapon systems can be considered along several continua. 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.

The 1970s and 1980s began to see more development of capabilities related to target identification, image discrimination, and target ranking or prioritization. These advancements are more than likely due to the technological advances made in sensor technologies in the 1970s, as well as in image-processing capabilities, through software development, microelectronics and microprocessor speeds, among others. What is more, the pursuit of long-range munitions required that they be able to direct themselves to particular targets, and, once there, identify those targets. Thus, strategic choices related to stand-off capabilities affected the acquisition and adoption of more self-mobile and self-directed weapons.

Today, with advances in machine learning, especially those related to image recognition and classification, there are movements to utilize these technologies in target recognition. Particularly, there is a desire to use advances in artificial intelligence to enable automatic target recognition so that the system can adapt and learn new targets when an adversary force changes tactics. Moreover, with growing capabilities to deny the manoeuvrability or use of stand-off weapons, militaries are also seeking to find new ways of utilizing miniaturization 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.

4. Areas of autonomous weapon development

There are potentially three areas to consider for autonomous-weapon system development: single platforms, combinations of legacy systems and modular systems.

Single platforms

Single-platform weapon systems or munitions, such as missiles, bombs, torpedoes or mines, are one potential area of autonomous weapon development. Such systems are better thought of as either a single platform (or swarm) with munitions on board, or as a single munition. The development of unitary autonomous weapons can be considered intentional. These are likely to be used in conjunction with other systems, but the systems can be thought of as "closed" or unitary. Maritime and air domains are the most likely areas in which these systems would be used, as there are fewer difficulties with obstacle avoidance.

Combinations of legacy systems

There is a likelihood that autonomous weapon systems will not appear first in the form of single platforms or single munitions. Rather, what is more likely is the combination of various legacy systems that enable a functionalist approach to autonomous weapon 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. Air, land and sea platforms may be combined in one 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, so that the actual choice of 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 weapon system (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.), the number of platforms in the collaborative operation, the geographical space within which the systems can function and the length of time that such systems can operate or extend operations by deploying further loitering submunitions, one could judge that, though no one single platform is an "autonomous weapon", the combination of multiple semi-autonomous systems yields an autonomous weapon system in a larger and functionalist sense.

Modular weapon systems

In contrast to the above scenario, in which existing platforms and munitions are combined to yield a functionally autonomous weapon system, with the modular approach to autonomous weapons, various parts of 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 weapon 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 parts, it being foreseeable that in certain combinations they may yield autonomous weapons. Such an approach could be domain-specific, such as the use of modular components with subsurface systems, or multi-domain, where components may fit on a variety of platforms or munitions in the air, on the ground and at sea.

SESSION 3: EMERGING TECHNOLOGY AND FUTURE AUTONOMOUS WEAPONS

Emerging technology and future autonomous weapons

Speaker's summary

Dr Ludovic Righetti, Max Planck Institute for Intelligent Systems, Germany

Trends in civilian robotics

The past few years have seen the emergence of several trends in civilian robotics. The technology necessary to create autonomous[2] cars, flying drones or underwater vehicles has existed for several years. It means that a car can drive autonomously (without any human intervention) with reliability. However, such cars are not yet available on the consumer market, mainly because of certification, reliability and liability issues. Who is responsible if an autonomous car is involved in an accident: the car manufacturer, the team that programmed the software driving the car or the car's owner? Autonomous vehicles make driving decisions without any human intervention, and therefore no human is directly responsible in the case of an accident. This exemplifies the difficulty of certifying the behaviour of an autonomous machine operating in a complex and unpredictable environment and ensuring reliability under those conditions.

Apart from autonomous driving or flying, complete autonomy in human environments that are constantly changing, and are not predictable, remains a great scientific challenge. First, a machine needs to use its sensors (cameras, Global Positioning System (GPS), etc.) to build a representation of the world that would allow it to make decisions (e.g. map its surroundings, detect people, recognize objects, etc.). Enabling a machine to understand its environment is extremely hard, and yet it is of major importance for any autonomous machine. In addition, algorithms that can make decisions based on this information are also very limited, and usually do not perform very well in complex and changing environments. The understanding of perception and decision-making algorithms that are scaled to complex and unpredictable environments remains a fundamental scientific issue in robotics research.

Therefore, many industrial or service applications of robotics are carried out either in environments of lower complexity (i.e. inside a factory where the environment is known in advance) or using supervised autonomy (i.e. a human operator still gives detailed instructions to a robot). For example, an operator can ask a robot to walk towards a goal, and the robot will control its balance and footsteps to walk in the desired direction without further intervention. If an unexpected event happens before the robot reaches the goal, the robot will stop and ask for further instructions. The US Defense Advanced Research Projects Agency (DARPA) Robotics Challenge of June 2015, which involved some of the most advanced robotics research laboratories in the world, was an example of supervised autonomy. In this case, robots had to achieve several tasks related to a disaster-response scenario, such as walking over complicated terrain, using a tool to break through a wall or climbing a ladder. For all these tasks, a remote operator was allowed to send commands to the robot to help it accomplish the tasks (e.g. by specifying good spots to place its feet or identifying a tool in an image). Even in the case of supervised autonomy, this challenge showed fundamental limitations in completing these tasks quickly and reliably.

[2] In the following, we use terms such as autonomy, decision-making or understanding. These terms refer to technical characteristics of machines and not philosophical concepts. For example, autonomy in robotics has nothing to do with free will, but relates to a machine's ability to accomplish complex tasks without human intervention.

How can we make sure that a machine will never fail?

From a technical point of view, it is impossible to guarantee that an autonomous machine will never fail, because it is impossible to enumerate all the possible combinations of events that might lead to a failure (people crossing the street, a car sensor failure, etc.). Can we at least guarantee that the worst-case failure will be limited, and know how often it might fail? This is a very difficult question to answer. For example, an autonomous car uses its sensors (cameras, 3D sensors, GPS, etc.) to build a representation of the world: where is the road? Are there any pedestrians trying to cross? Is there a traffic light? It also uses other sources of prior knowledge, such as a map of the area where it can locate itself, potentially containing an indication of the location of traffic lights or current construction areas. After combining these pieces of information, the car algorithm (i.e. the software program) makes a driving decision: break, accelerate, turn, etc. Since the algorithm that makes the decision is based on this constructed representation, it is very hard to predict what will happen in every possible situation. What will happen if one sensor does not work very well, if someone tries to trick the perception system by jumping around the car, or if the car is in a situation that has never been seen before? While it is possible to test the machine in many situations, it is impossible to test for every possible occurrence in an unpredictable environment. While there are no absolute guarantees, methods are being developed to provide at least statistical information about the likelihood of failure and guarantees in relation to the worst-case scenarios. Nevertheless, as machines become more complex and more autonomous, providing such guarantees becomes harder. For example, one can show that an autonomous car is working well by driving it many thousands of miles under various weather and traffic conditions. Therefore, it is possible to say that there is a high probability that the car will keep working well, but it is impossible to guarantee that it will never cause an accident.

What is machine learning?

Another trend that has been publicized recently by the media is the progress made in machine learning and its consequences. Machine learning is a field of science mainly concerned with the problem of finding statistical relationships in data. Machine-learning algorithms are increasingly being used in robotics and other engineering fields. By exploiting data generated from real-world examples, one can create algorithms capable of very high performance. For example, detecting a cat in an image is best done using machine learning. It is important to understand that machine learning, despite what it suggests, does not correspond to learning in the sense understood for humans. Machine learning is roughly divided into three categories[3]: supervised learning, reinforcement learning and unsupervised learning.

Supervised learning uses a set of examples with a label informing the algorithm of the expected output. For example, if we have a large dataset containing images of cats and images without cats, we can use machine learning to create a classifier that will be able, after learning, to decide whether there is a cat in the picture (or, more frequently, it will give us the probability that the image contains a cat). Deep learning is a supervised learning technique based on artificial neural networks that was invented decades ago, but that has become extremely successful recently owing to improvements in computing power. Neural networks, while inspired by the connectivity of the brain, have nothing to do with a human brain. They are just a convenient way to represent a mathematical function by using many simple units (the artificial “neurons”) connected together. Each unit computes a number based on its inputs. The output of the neural network will be something like the sum of the output of all these units.[4] Deep learning consists of many layers of these neural networks connected together, and is very effective in extracting the statistical relationship between inputs and

[3] This follows the description by Yann LeCun, a leader in deep-learning research, https://interstices.info/jcms/p_89081/les-enjeux-de-la-recherche-en-intelligence-artificielle (in French).
[4] Note that in practice it is a bit more complicated than just a summation, but the idea remains the same.

outputs using massive amounts of data. Its biggest successes so far are in computer vision and language processing. For example, these neural networks can be trained to recognize objects in a picture (such as the cat in our previous example) with a very high degree of accuracy. While it represents extraordinary technology and allows very complicated problems to be solved, deep learning is very far from any form of intelligence.

In reinforcement learning, algorithms learn how to choose between a set of actions to accomplish a task that will maximize some reward. The reward is a mathematical function that computes a score depending on how well the task was completed (i.e. the higher the score, the better). Through trial and error, by looking at how the reward changes, the algorithm is able to find actions that will increase the reward in future decisions. For example, reinforcement learning was used, in conjunction with deep learning, to create the program AlphaGo that defeated a professional human player at the game of Go recently. Given an image of the board game, the program learned how to decide where to put the next stone on the board in order to increase its chances of victory (the reward). The program had to choose between a limited number of actions, i.e. a position on the board, from an image of the game. Such algorithms can only make a decision within a set of possible actions (e.g. the position of the stone in our example) and cannot come up with new actions. For example, the software will not decide suddenly that it wants to play chess. In this case again, learning algorithms are not doing anything related to human intelligence: the algorithms can become really good at playing a well-defined game, but cannot decide to play another game. A good analogy for understanding how this differs from human intelligence is the mechanical excavator, which is much better than humans at moving a large amount of soil, in a similar way that a program can be much better at playing chess than a human. But that does not make either the excavator or the program intelligent, because they cannot do anything else.

Finally, “unsupervised learning” refers to the problem of designing algorithms which can learn by themselves without any external goal (either a list of labelled examples or rewards) and which would be able to come up with their own goals. Many people believe that it is the key to creating really intelligent machines. However, it is fair to say that so far machine-learning research has not provided the technology capable of solving this problem, and no one knows if it is even possible to do so.

The problem of predictability

When using machine learning, since we extract statistical relationships in data, there is an issue related to predictability. What happens if the algorithm is given input data that are vastly different from what it has encountered before? It is very difficult to predict this outcome reliably if the system is complicated. In many cases, this is not a problem (e.g. sometimes the algorithm shows a picture of a dog instead of a cat), but we see that it can be a problem when the algorithm’s output is used to make a safety-critical decision (e.g. is there a pedestrian crossing the road?). Due to the nature of the algorithms, it is not possible to guarantee with 100% certainty that it will always work, and it is usually only possible to give probabilities of success and failure (e.g. the algorithm detects a cat in 99% of the images containing a cat and detects a cat in 2% of the images not containing one).

As we have seen above, the reliability and robustness to failure of a machine become more challenging as autonomy increases. In addition, subcomponents using machine-learning algorithms are being used increasingly in robotics and in computer-science fields in general. For example, it is now standard for computer vision and image recognition to use deep learning to train image classifiers. Therefore, an additional source of unpredictability is added to complex robots, and it becomes harder to provide strong guarantees of the behaviour of these machines. This can be acceptable for an autonomous car which is extensively road-tested, which can be provided with several fail-safe modes (e.g. give control back to the human) and for which the worst-case outcome will be causing an accident no worse than

what humans would cause. But we can see why this could be problematic in more critical situations, where a failure can have more disastrous consequences for machines which operate at faster speeds and on larger scales.

While predictability can be an issue, it is important to stress that any algorithm used in a robot has a well-defined scope of behaviour. First, machines make decisions following an algorithm, i.e. machines just do what they are programmed to do, whatever the complexity of the program,[5] and so there is no unpredictability in terms of whether the machine will decide to do something it was not programmed for. It will never do something it was not programmed for, and this is also true when using machine learning. The unpredictability comes from the uncertainty of the environments, the complexity of the algorithms and potential unexpected failures. But this can be statistically quantified (as for an autonomous car), and it might also be possible to give bounds for worse-case behaviours, despite this being very complex to determine.

[5] It is important to emphasize that when machines make decisions, this has to be understood from an algorithmic point of view: they just follow the algorithm, and the output of the algorithm is what we call the decision. Machines have no conscience or related higher-level characteristics associated with human intelligence.

SESSION 4: LEGAL AND ETHICAL IMPLICATIONS OF INCREASING AUTONOMY

Legal issues concerning autonomous weapon systems

Speaker's summary

Col. Zhang Xinli, Ministry of Defence, China

As noted in the ICRC's background paper, there is no internationally agreed definition of autonomous weapon systems. According to the ICRC's working definition, "autonomous weapon systems" are weapons which can independently select and attack targets, i.e. with autonomy in the "critical functions" of acquiring, tracking, selecting and attacking targets. Based on this definition, this presentation will discuss the challenges posed by autonomous weapon systems to international humanitarian law (IHL) and their legality under international law as a whole.

Challenges of autonomous weapon systems to IHL

Autonomous weapon systems, like other new weapons, should be reviewed in the light of IHL rules. Under IHL, including the Geneva Conventions of 1949 and their two Additional Protocols of 1977, there are some fundamental principles concerning the use of means or methods of warfare. These are the principle of distinction, which provides that means of warfare shall discriminate between civilians and combatants, and between military objectives and civilian objects; the principle of proportionality, which requires that the incidental civilian casualties expected from an attack on a military target not be excessive when weighed against the anticipated concrete and direct military advantage; and the principle of restriction, which restricts the use of some cruel weapons in armed conflict. The purpose of these principles is to minimize the suffering caused by armed conflict while not impeding military efficiency. There are some concerns regarding the ability of autonomous weapon systems to comply with some of these principles and related rules.

First of all, whether autonomous weapon systems have the ability to distinguish legitimate targets is questionable. Secondly, autonomous weapon systems pose challenges to the principle of proportionality. Thirdly, it is difficult to determine individual responsibility. One of the important measures to protect the victims of armed conflict is to investigate individual criminal responsibility for grave violations of IHL. Autonomous weapon systems have no sense of ethics; it would make little sense to attribute responsibility for violations to a computer or other machine. Thus, it is difficult to determine who would be accountable for violations of IHL committed by an autonomous weapon system.

Finally, autonomous weapon systems pose a challenge to the peaceful resolution of international disputes. They may decrease the costs of waging war for those countries with technical advantages. Such countries may tend to use force instead of peaceful means to settle international disputes. As a result, civilians and soldiers from other less technically advanced countries may bear a greater loss. This will undoubtedly cause a catastrophe in humanitarian terms.

The legality of autonomous weapon systems needs more discussion

Although autonomous weapon systems pose a series of challenges to IHL, it is hard to draw definite conclusions that autonomous weapon systems are inherently illegal.

One of the reasons for this is that the understanding of autonomous weapon systems is still in the realm of imagination. According to the definition of autonomous weapon systems used by the ICRC, "fully" autonomous weapon systems are still at the research stage. Truly autonomous weapon systems have not yet appeared, let alone been deployed in armed conflict. So the current study and discussion is based on possibilities and assumptions, which makes it hard to avoid bias. Another factor is that we should take a more comprehensive view of the challenges raised by autonomous weapon systems to the rules of IHL.

Last but not least, for the time being, there is no specific international treaty banning autonomous weapons. According to the existing international conventions, the use of weapons which would cause excessive damage and suffering, such as toxic, chemical and biological weapons and certain conventional weapons, is forbidden or restricted. It is difficult to classify autonomous weapons in this category. In addition, the relevant international conventions also ban the use of indiscriminate means and methods of warfare, as well as those means and methods that would harm the environment. From the point of view of the existing research, autonomous weapon systems are not designed to damage the environment. As for indiscriminate means and methods of warfare in relation to the principles of distinction and proportionality, this was raised in the previous section.

Conclusion and prospects

Generally speaking, the current international discussions on autonomous weapon systems are at a preliminary stage. There are many aspects of such systems that warrant further in-depth study and analysis, including their definition, whether the existing international legal framework is adequate to regulate these emerging weapon systems, and their potential impact on global security and stability. At this stage, it is still too early to reach any conclusions on the above questions, or on whether IHL is the only criterion we should consider when judging a new weapon system. This expert meeting organized by the ICRC provides a good opportunity for officials and experts from different countries to engage in meaningful and necessary discussions. In my view, in order for international society to conduct more substantive discussions that might eventually launch a result-oriented international process, we could proceed as follows:

Firstly, attention should be focused on the definition of autonomous weapon systems. Though reaching a universally accepted definition is by no means an easy task, we should be aware that a clear definition is the foundation of further meaningful discussion, and that a precise definition in legal terms is a precondition for discussions on the legality of such systems, as well as for the prohibition of their development and use. Most of the current proposals adopt a technical approach, namely distinguishing autonomous weapon systems from other weapon systems based on their components, key functions and level of human control, or the context where the weapon is used. These approaches offer useful insights. Taking into consideration the current level of artificial intelligence, the ultimate decision to use a weapon still lies in human hands, and the systems we are talking about should be considered future weapons with sufficiently high artificial intelligence to be used autonomously. A technical threshold could be set for distinguishing autonomous weapon systems from other weapon systems. When defining autonomous weapon systems, we could combine a description of their key technical features with references to specific weapon systems. A feasible definition should capture the main technical features while taking into consideration possible future developments in autonomous weapon systems.

Secondly, national legal reviews should be viewed objectively. Such reviews play a positive role in ensuring compliance with IHL. As required by Article 36 of Additional Protocol I to the Geneva Conventions, States should conduct domestic legal reviews when developing a new weapon, and countries should take the necessary measures, pursuant to their national laws and regulations, to ensure compliance with their international obligations. However, such national reviews are not enough to ensure the legality of a new weapon, because certain questions – for example, to what degree can a simulated environment match the complex and dynamic environment in the field, and to what degree could a unilateral review withstand outside supervision, thus ensuring its effectiveness – are subject to further discussion. The international community should be clearly cognizant that domestic review, if overemphasized, could provide a legal pretext for some future weapon system that should not have been developed in the first place.

Thirdly, the development of an international instrument on the prohibition or limitation of autonomous weapon systems is a long and complicated process. Because of the complexity of issues concerning autonomous weapon systems, the close relationship between military and civilian uses of artificial intelligence, and the implications it could have for the development of future technology, such a procedure should be initiated in the context of an in-depth and full discussion, and of consensus on key aspects of autonomous weapon systems. When undertaking this arduous task, the international community must strive to keep a balance between addressing humanitarian concerns and legitimate national-security concerns, so as to attract as many countries as possible. At the same time, such an instrument should not unduly constrain the development of civilian technologies which could provide impetus to social development, nor should it set a new technical barrier to the large number of developing countries that are not currently actively involved in the process.

Fourthly, more outreach to developing countries is needed so as to ensure wide and equitable participation. The international discussion on autonomous weapon systems has been going on for a few years, but only a small number of countries have voiced their views. The vast majority of developing countries are silent on this topic. They are either not aware of its importance or not interested in the discussion. With a view to developing a widely accepted international instrument, more developing countries should be encouraged to join in this process. In this regard, international cooperation and assistance are needed to raise their awareness of the topic.

Autonomous weapon systems and IHL compliance

Speaker's summary

Dr Gilles Giacca, Legal Adviser, Arms Unit, Legal Division, ICRC

Note: For a summary of the issues raised in this presentation, see Section 5 of the ICRC's background paper in Part III of this report.

Autonomous weapon systems and the alleged responsibility gap

Speaker's summary

Prof. Paola Gaeta, The Graduate Institute, Switzerland

This presentation aims to clarify whether there is an accountability gap for violations of international humanitarian law (IHL) by autonomous weapon systems. This presentation argues that such a gap does not exist with regard to State responsibility. Regarding criminal liability, however, the subjective element (*mens rea*) could be hard to fulfil in some situations. This is especially true for the Statute of the International Criminal Court (ICC), which contains a narrower definition of the *mens rea* than does customary law. However, before national courts and tribunals applying customary international law, it can be possible to establish criminal accountability via indirect intent.

General difficulties concerning the autonomy of machines and criminal law

The following example illustrates the difficulties posed by autonomous systems – not necessarily only by weapon systems – with respect to criminal accountability. It is based on facts. A group of Swiss artists created a program called "Random Darknet Shopper," which was programmed to spend a certain sum on the darknet on a daily basis. In the end, it purchased 16,000 items, including illegal goods, such as ten Ecstasy pills, a fake Hungarian passport and a fake Louis Vuitton handbag. When we look for the person who is criminally responsible in this scenario, there are three options: the programmer, the user or the robot itself.

The current debate on the alleged responsibility or accountability gap with regard to autonomous weapon systems revolves around the answers to the aforementioned questions. In the academic literature, all three options (holding the programmer, the user or the machine itself accountable) have been proposed.

This presentation focuses on autonomous weapon systems carrying out targeting decisions on the battlefield without human interference ("human out of the loop"). It has been argued that it would be unfair to make the human out of the loop responsible for any violation of IHL amounting to a war crime "committed" by the machine. Such lack of accountability is said to increase the risks of unlawful attacks with killer robots.

It is doubtful whether fully autonomous weapon systems will ever exist. Let us assume, however, that a machine operating completely independently from humans commits a violation of IHL. Even though it has been suggested at times that the machine itself should be held accountable, this is not possible under criminal law, which presupposes human actions. The programmer could be responsible, but often his or her involvement is quite distant from the execution of the actual attack. This leaves the commander as the closest human link to the attack. The chain of command would even be shorter than in the usual scenario of soldiers on the battlefield. Could the commander be held responsible?

In this case, the causality requirement of *conditio sine qua non* would not be more difficult to meet than for a human subordinate; the same goes for the other objective elements of a crime (*actus reus*). However, the *mens rea* will be hard to prove. In most cases there will be no direct intent to use the autonomous weapon system to commit a war crime, but only an "acceptance of the risk" that the machine may take the wrong targeting decision. The question remains: is this acceptance sufficient in and of itself?

Criminal responsibility: The issue of *mens rea* and war crimes

At the ICC the standard for *mens rea* is high. Article 30 of the ICC Statute and relevant war crimes provisions (such as those on targeting civilians) require direct intent, although there is no need to prove that civilians were actually killed. It is a crime of conduct as opposed to a crime of result. This means that it is not possible to conclude that the *mens rea* is fulfilled unless the officer *intended* to commit a violation of IHL or at least knew with *certainty* that such a violation would occur.

But despite this gap, there remains another option for criminal liability. Article 85 of Additional Protocol I to the Geneva Conventions (AP I) on "grave breaches" requires "wilful" targeting of civilians. The requirement of wilfulness was interpreted by the International Criminal Tribunal for the former Yugoslavia (ICTY) as including "indirect intent". This means that the acceptance of the risk that a certain behaviour *might* result in a certain outcome is sufficient to fulfil the element of *mens rea*. The ICRC commentary on AP I concurs with this interpretation of "wilful."

In short, this means that States party to AP I remain bound by their obligations under it. Article 85 of AP I lists grave breaches which need to be criminalized in national legislation. All States Parties thus remain bound by the lower threshold of indirect intent contained in Article 85 of AP I and need to legislate accordingly. From this angle, the accountability gap seems less wide, since war criminals can be tried by national courts. Furthermore, it is accepted under customary international law that indirect intent suffices for the commission of a crime, unless otherwise stated. Thus, even an international tribunal applying customary international law would face fewer challenges with regard to *mens rea*, and could apply the lower standard.

State responsibility

Finally, one should not forget that criminal responsibility is not the only way to establish accountability for violations of IHL. The framework of State responsibility can equally serve this purpose.

The great advantage of State responsibility is that, in contrast to criminal law, it does not require a mental element. It is sufficient for a violation of international law to be objectively attributable to a State, for example because it was committed by a person acting on the State's behalf. The State in question would be responsible for the violation, unless it successfully invokes *force majeure*. The threshold of *force majeure*, however, is very high. An ordinary malfunction of an autonomous weapon system would not suffice, although a completely unexpected incident against which no reasonable precautions could have been taken would qualify. However, the burden of proof rests with the State.

The additional advantage of State responsibility is the State's obligation to make full reparation to the victims, including compensation. In this sense, the State responsibility framework is even more effective than international criminal law, where the idea of compensation for the victims exists only at the ICC (in a more rudimentary way). Bearing this in mind, State responsibility could thus have a considerable deterrent effect on States and would give them an incentive to make sure that the autonomous weapon systems deployed comply with IHL.

SESSION 5: HUMAN CONTROL

Meaningful human control over individual attacks

Speaker's summary

Mr Richard Moyes, *Article 36*, UK

Introduction

"Meaningful human control over individual attacks" is a phrase coined by the non-governmental organization *Article 36* to express the core element that is challenged by the movement towards greater autonomy in weapon systems. It is a policy formulation that has been picked up and used in different ways: in publications by various individuals and organizations, in statements at review conferences of the States party to the UN Convention on Certain Conventional Weapons (CCW), in the open letter from artificial intelligence practitioners organized 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 element that autonomy removes, and therefore describing this element is a necessary starting point if we are to evaluate whether current or future technologies challenge it. This is particularly important if we are to have a coherent policy conversation about diverse and often hypothetical future technologies. It is also a starting point for policy that is arguably more open to engagement by diverse parties who might have different expectations of the advantages that future developments in autonomous weapon systems might provide to them.

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 issues fall under the key elements of human control: for example, the need for "predictable" technology, the need for human judgement to be applied in the use of force, and the need for accountability, which we will look at later. Furthermore, without a normative requirement regarding human control, the legal framework itself is open to divergent and progressively broader interpretations that may render human application of the law meaningless.

Recognizing the need for human control in some form

At its most basic level, the requirement of meaningful human control develops from two premises:

1. that a machine applying violent 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 should be assessed against a common standard and so necessarily 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 representatives' anxiety over policy formulations not initiated by States. Such responses, however, miss the point. There are other words that could be used instead of "meaningful", e.g. appropriate, effective, sufficient, or necessary. Any one of these terms leaves open the same crucial question: how will the international community delineate the main 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), it derives from an overarching principle rather than being outcome-driven (e.g. effective or 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 delineating the elements of human control that should be considered necessary in the use of force.

Situating human control in the legal framework

Article 36 has called on States, in the context of discussions on autonomous weapon systems in armed conflict, to recognize the need for "meaningful human control over individual attacks." By 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 recognize that IHL is not the only legal framework relevant to autonomous weapon systems, nor are legal frameworks the only basis for assessing whether the further development of such technologies is appropriate or advisable. However, the relationship between human control, autonomous weapon systems and IHL are given particular focus here.

Human beings as addressees of the law

When discussing autonomous weapon systems, however complex, *Article 36* orients to these systems as "machines". The discussion of this issue is prone to 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 autonomous weapon systems will "be able to apply legal rules", or "to follow the law". Machines don't 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 human commanders anticipate in their assessment of the legality of a planned attack. Prof. 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".[6]

[6] ICRC, *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*, Report of an expert meeting held in Geneva, Switzerland on 26–28 March 2014, November 2014, p. 41.

Human judgement in relation to "attacks": part of the structure of IHL

If human beings are the addressees of the law, whether collectively or individually, then there are certain boundaries of machine operation that the law implies in relation to humans. The term "attacks" in IHL designates a unit of military action, and it is to individual "attacks" that certain legal judgements must be applied. So attacks are part of the structure of the law.

For example, Article 57 of AP I 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. Therefore, it is humans who shall apply these legal rules, including verifying the objective, choosing the means and method 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; otherwise, it is indiscriminate (Article 51.4(a)). 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), and that in the application of the legal rules, the concrete and direct military advantage must be assessed by humans who plan and decide upon an attack (Article 51.5(b) and Article 57.2(a)(i) and (iii)). 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).

These rules imply that a machine cannot identify and attack a military objective without human legal judgement and control being applied in connection with an attack on that specific military objective at that time (control being necessary in some form to act on the legal judgement that is required). Arguing that this capacity can be programmed into the machine is an abrogation of human agency with respect to the law, 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 autonomous weapon systems 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 judgements at certain points.

However, this is not to argue that the law straightforwardly implies a very narrow constraint on what an autonomous weapon system might do under its existing terms. Nor is it suggesting that existing law alone represents a sufficient basis for managing such weapon systems. It is simply to point out that the existing legal structure (human judgement being required with regard to "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).

Conceptualizing "an attack"

While an assumption of human legal judgement in relation to individual attacks is seen in the structure of the law, it is also recognized 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 conceptualized – from the local tactical level, through the operational level, to the broad strategic level. If "attacks" were not conceptualized and subject to legal judgement at the tactical level, but only at the broad strategic level, then a large operation may be determined to be permissible (on the basis of broad anticipated outcomes) while containing multiple individual actions that would in

themselves be violations of the law. Clearly, for the law to function meaningfully, there need to be legal judgements and accountability for actions at the most local level.

Recognition that human legal engagement must occur with each attack means that a machine cannot proceed from one attack to another without such human legal judgement being applied in each case, and without the 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, under the law, an attack is carried out against a specific military objective that has been subject to human assessment in the circumstances prevailing at the time, it follows that a machine cannot set its own military objective without human authorization based on a human legal judgement.

Preventing an expansion of the concept of "an attack"

Our starting point in this discussion was concern that greater autonomy in weapon systems may result in human control not being meaningful. Based on the above analysis 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 weapon systems might produce an expansion of the concept of "an attack" away from the granularity of the tactical level towards the operational and strategic levels. That is to say, there is a risk of autonomous weapon systems being used in "attacks" which, in their overly broad spatial, temporal or conceptual boundaries, go significantly beyond the units of military action over which specific legal judgement would currently be expected to be applied.

A more specific legal assessment – in other words, a legal assessment of specific events that are expected to occur over a shorter period of time and within a narrower area – makes it possible more accurately to assess specific risks to the civilian population and therefore to enhance protection of civilians. Furthermore, allowing greater autonomy to facilitate progressively broader interpretations of what constitutes an attack would have a corrosive effect on 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 the permissibility of a weapon system are being undermined by the development of that weapon system itself, how can the review process based solely on those tests remain meaningful?

By asserting the need for meaningful human control over attacks in the context of autonomous weapon systems, States would be affirming a principle intended to protect the structure of the law as a framework for the application of wider moral principles. Moving the debate onward to delineate the elements needed for human control to be meaningful would foster a normative understanding that should pull towards greater specificity in legal assessment, rather than greater generalization.

Key elements of human control

Thus, as outlined in 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 lays out key elements through which human control can be understood to be applied in the use of weapon systems. These elements are not simply about technological characteristics; they recognize 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 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); and
- transparent (practical users can understand how it works).

However the technology is to be used, it can be designed and manufactured with certain characteristics that have a bearing upon the subsequent capacity for human control. A technology that is by design unpredictable, unreliable and non-transparent is necessarily more difficult for a human to control in a given situation of use.

Providing accurate information for the user on the outcome sought, the technology and the context of use

Human control in the use of a given technology is thus based on those who plan and decide upon an attack having certain information. Control in the use of a weapon system can be understood as a mechanism for achieving the commander's intent. So information on the objective sought – and among other things, on the unintended consequences that a commander wishes to avoid – is an important starting point. 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 weapon system will identify as a target object – i.e. 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 whether the force consists of a heavy explosive weapon with a large blast and fragmentation radius, or whether force will be applied quite narrowly, e.g. through an explosively formed projectile with no fragmentation effects.

Predictability is an important concept, in that it provides a link between the commander's intent and the likelihood of outcomes matching 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 that enhances our understanding of the context of use, including the presence of civilians and civilian objects, for example, is very significant.

Of course we may not achieve complete predictability. Already, in the use of weapons, commanders 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 of its operation. For any given environment, it follows logically that a larger area and longer duration of independent operation by a technology result in reduced predictability and thus reduced human control.

It is recognized that different environments have different general characteristics, with the land, air and sea presenting different levels of complexity. This may mean that a large area of operation at sea may still facilitate better contextual understanding than a smaller area on land. However, given environments of equal complexity, a larger area and longer time of operation still 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 might be because the commander's intent has changed from the time when the attack was initiated.

From an understanding of the technology and the context in which it will operate, a

commander should be 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 the understanding of these three elements must coalesce to a degree where an informed judgement can be made.

Timely human judgement and action, and the potential for timely intervention

Based on the information on the outcome sought, the technology and the context, we need humans to apply their judgement, as implied by the legal analysis above, and to choose to activate the technology. This point of human engagement ties together the systems of information upon which judgements are made, but also provides a primary reference point 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 judgement and action are 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 – about the context, for example – also degrade over time. For a system that may operate over a longer period, a clear capacity for timely intervention (e.g. to stop the independent operation of a system) will be necessary if it is not to operate outside the framework of 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 which produces and maintains the technology and which 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 extent of human control itself to be assessed as sufficient. Where some have asserted that the existing legal framework provides the answers needed for evaluating autonomous weapon systems, these tests suggest that this is not straightforwardly the case.

It is not clear, for example, what level of information about the context in which a weapon will be used is considered sufficient to provide a basis for an informed legal judgement. If a weapon 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 bounded geographical area of which a commander had knowledge. However, if the area in 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 what can reasonably be considered an attack. The law alone does not provide an answer to these questions that resolves 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 weapon systems.

While 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 recognize 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 weapon systems.

Working definitions: facilitating discussion within the framework of the CCW

The most direct way to establish such a discussion within the framework of the CCW is to adopt an approach to working definitions that is based on the recognition that certain forms of human control over the use of force are required, and that systems operating outside such control should not be considered acceptable. That would most straightforwardly be facilitated by adopting a working definition of "lethal autonomous weapon systems" that is based on their being "weapon systems operating with elements of autonomy and without the necessary forms of human control". In such an approach, the concept of weapon 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 paves the way for 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.

Human control in the targeting process

Speaker's summary

Ms Merel Ekelhof, VU University Amsterdam, Netherlands

Overview

This presentation aims to provide one way of looking at the use of autonomous weapon systems. I will offer an analysis of the targeting process as it relates to the debate on such weapon systems. As these systems do not operate in a vacuum, it is relevant to provide some context concerning targeting and the use of weapon systems by the military. To clarify the concept of "meaningful human control", I will first briefly explain the definition used. Then I will continue with an illustration of the targeting process in order to show how human control is currently exercised over weapon systems with autonomous functions. By providing this context, I intend to guide the thinking about the use of autonomous weapon systems and present a way of looking at the concept of meaningful human control.

Working definition

Through the many debates taking place, different approaches to autonomous weapon systems are being shared and different terminology is being used. Consequently, the current debate relies on language too imprecise and indefinite to clearly define autonomous weapons. It is therefore important for the definitions used to be explained beforehand. The working definition that will be used throughout this presentation to describe an autonomous weapon system follows the definition proposed by the ICRC:

Any weapon system with autonomy in its critical functions. That is, a weapon system that can select and attack targets without human intervention.[7]

Although it is sometimes argued that autonomous weapon systems do not yet exist, this definition does include some existing weapon systems with autonomy in the critical functions of selecting and attacking targets. Examples have been given during earlier sessions on missile- and rocket-defence weapons, vehicle "active-protection" weapons, loitering munitions and torpedoes. It could be useful 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 of greater autonomy are being pushed.

The targeting process

The *loop* has become a very familiar term in the debate about the use of autonomous weapons. Generally, the loop is explained as having three categories: weapons with a human "in the loop", weapons with a human "on the loop", and weapons with a human "out of the loop". "Human in the loop" is regularly explained as the capability of a machine to take some action, but then stop and wait for a human to take a positive action before it continues. Then, there is the phrase "human on the loop", meaning that humans have supervisory control and only intervene to stop a machine's operation. The phrase "human out of the loop" is often used to describe autonomous weapons, because it would mean that the machine will take some action and the human cannot intervene. [8]

[7] ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, Report to the 32nd International Conference of the Red Cross and Red Crescent, Geneva, held on 8–10 December 2015, October 2015, pp. 44–47.
[8] Statement by Prof. Paul Scharre, Center for a New American Security, at the session on technical issues, CCW Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, 13 April 2015; Human Rights Watch and

The advantage of using the loop metaphor to describe autonomy in weapon systems is that it focuses on the human-machine interface. It seems to be a useful device, because people can potentially more easily relate to their role as a human operator or supervisor than conceive of something as complex and debatable as autonomy. Nevertheless, it is not always clear what is meant by "loop". According to Peter Singer, there is a movement afoot to redefine the meaning of having a human "in the loop".[9] Ray Kurzweil argues that "in the loop" is becoming no more than "a political description".[10] And 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.[11] Clearly, what is meant by having a human in, on or out of the loop is not always straightforward. I propose to explain the loop as the targeting process that is used by the military to plan, execute and assess military missions.

The term *targeting* is often associated with the actual use of force, i.e. a lethal attack or kinetic action, such as firing a weapon at a target. However, the targeting process entails more than the actual kinetic action; there is, as the name implies, an entire process or decision-making cycle that precedes or surrounds this moment. NATO's targeting process serves as an example of how weapons are used and how humans can exercise control over increasingly autonomous weapon systems.[12]

Targeting is an iterative process which aims to achieve mission objectives in accordance with the applicable law and rules of engagement through the thorough and careful execution of six phases. NATO explains the phases as follows:

1. Commander's objectives and guidance are formulated during which the commander must clearly identify what to accomplish, under what circumstances and within which parameters;

2. Targets are developed, nominated, validated and prioritized. 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.[13] Both the principle of distinction and issues related to collateral damage play a role;

3. Capabilities are analysed to assess what methods and means are available and most appropriate to generate the desired effects;

4. Capabilities are matched to the targets. This phase integrates output from phase 3 with any further operational considerations;

5. The assigned unit will take steps similar to those in phases 1 to 4, but on a more detailed, tactical level. And, importantly, there is force execution, during which the weapon is activated, launched, fired or used; and

6. Combat is assessed to determine whether the desired effects have been achieved. This feeds back into phase 1, and the goals and tasks can be adjusted accordingly.[14]

Harvard Law School International Human Rights Clinic, *Losing Humanity – The Case against Killer Robots*, https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots, 2012; Defense Science Board, *Task Force Report: The Role of Autonomy in DoD Systems*, July 2012; M.N. Schmitt and J.S. Thurnher, "'Out of the Loop': Autonomous Weapon Systems and the Law of Armed Conflict", *Harvard National Security Journal*, Vol. 4, No. 2, pp. 231–281; ICRC, *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects,* Report of an expert meeting held in Geneva, Switzerland on 26–28 March 2014, November 2014, p. 41.

[9] P.W. Singer, *Wired for War*, The Penguin Group, New York, p. 125.

[10] *Ibid.*

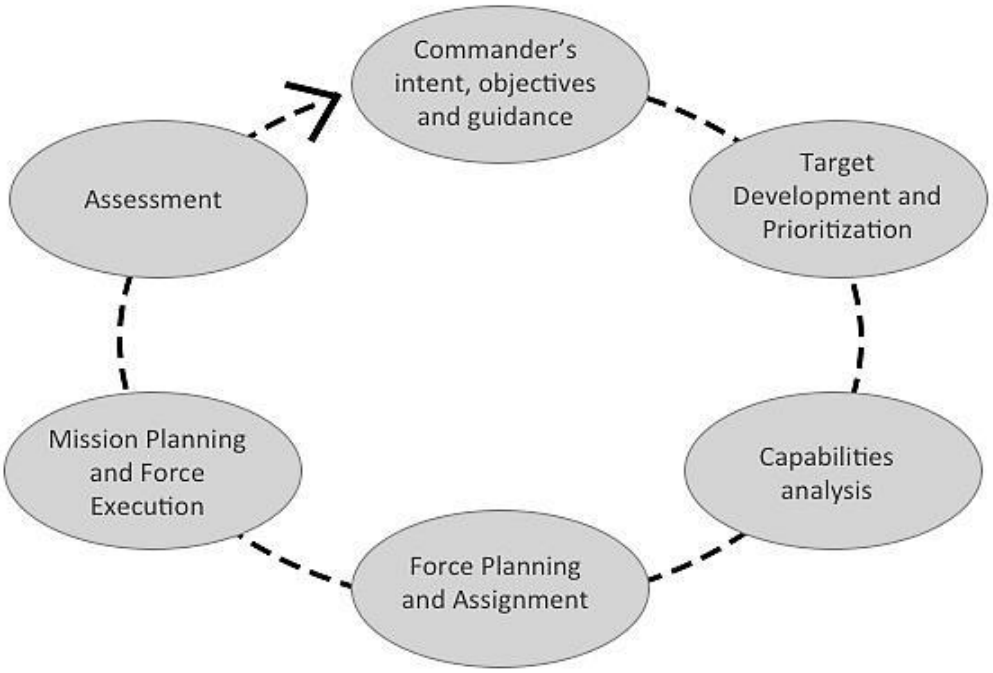
[11] M.C. Marra and S.K. McNeil, "Understanding 'The Loop': Regulating the Next Generation of War Machines", *Harvard Journal of Law & Public Policy*, Vol. 36, No. 3, pp. 1139–1185.

[12] NATO Allied Joint Publication (AJP)-3.9, *Allied Joint Doctrine for Joint Targeting*, May 2008, pp. 2-1–2.4.

[13] Specific IHL rules cannot be categorized 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. Joint Committee on AWS, *Autonomous Weapon Systems: The Need for Meaningful Human Control*, Netherlands Advisory Council on International Affairs (AIV) No. 97/Netherlands Advisory Committee on Issues of Public International Law (CAVV) No. 26, October 2015, http://aiv-advice.nl/8gr/ http://www.cavv-advies.nl/3bz/home.html.

[14] *Ibid.*

The following diagram is an illustrative example of the different steps in a targeting process. It is an oversimplification. Targeting 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 in which weapon systems with autonomy in their critical functions operate.



Human control in the targeting process

As mentioned previously, an autonomous weapon is described as a weapon that can select and attack targets without human intervention, of which some current examples exist. These include 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.[15] This is the case, for example, with the Israeli Harpy, which is programmed to select and engage hostile radar signals in a predefined area. After activation, humans can no longer intervene in the process of target selection and attack. However, that does not mean that humans are not in control of the autonomous weapon system. 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 they set overall goals, gather intelligence, select and develop targets, identify the most suitable weapon, and decide in what circumstances and under what preconditions to employ a particular weapon. Thus, even though an autonomous weapon system selects and attacks a target in phase 5, it is not truly autonomous in the overall targeting process. It is through this process that humans can remain in control of an autonomous weapon system's actions on the battlefield, even though there is no direct human control over the system's critical functions of target selection and attack.

Within the targeting process, humans can exercise control in different ways. Humans can assign operational constraints by programming a predefined geographical area to which the

[15] M. Roorda, "NATO's Targeting Process: Ensuring Human Control Over (and Lawful Use of) 'Autonomous' Weapons", in A. Williams, P. Scharre (eds), *NATO Headquarters Supreme Allied Commander Transformation Publication on Autonomous Systems,* p. 16.

autonomous weapon system has to limit its operation, and by limiting the time within which the weapon system is allowed to operate without direct human control. In addition, humans can set parameters for their judgement by, for example, requiring humans 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.[16] The targeting process provides opportunities for humans to actually exercise these key elements or forms of (meaningful) human control in an organized and structured manner.

Future technologies

In this presentation, I decided to look at the targeting process and analyse the use of *current* weapon systems with autonomy in their critical functions to see how meaningful human control is understood and implemented today. But that is not where the analysis should end. Because of 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 with no 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.

Conclusion

A lot of the debate on autonomous weapon systems fails to take into account the military targeting process. It is through this process of planning, execution and assessment that humans can remain in control of an autonomous weapon system's actions, even though there is no direct human control over the system's critical functions of target selection and attack. Hence, it is important to point out such existing processes when considering the use of these weapon systems.

The aim of this presentation was not only to illustrate human control in the targeting process. That process also served as an example of decision-making which deserves specific attention when we are thinking about the concept of meaningful human control. It is not only the AWS itself and the human responsibilities that deserve attention, but also the decision-making process in which they interact. The targeting process is one in which humans and machines collaborate. It illustrates the context of *use* of these systems and offers one way of looking at meaningful human control. But meaningful human control is a concept that also plays a role in weapon *design* and *development*. Therefore, other decision-making processes – such as (legal) review processes – should also have a part in defining the concept of meaningful human control. In each of these processes, humans have responsibilities. Increasingly autonomous technologies could influence and perhaps lead to a redistribution of certain responsibilities within these processes. We should consider this and make sure that meaningful human control remains present in all forms of decision-making in which humans and autonomous weapon systems interact..

[16] M. Horowitz and P. Scharre, *Meaningful Human Control in Weapon Systems: A Primer*, Working Paper, Center for a New American Security, http://www.cnas.org/human-control-in-weapon-systems, p. 4.

SESSION 6: ADDRESSING THE CHALLENGES RAISED BY INCREASING AUTONOMY

Lethal Autonomous Weapon Systems (LAWS)

Speaker's summary

Lt Col. John Stroud-Turp, Ministry of Defence, UK

The debate concerning lethal autonomous weapons, or LAWS, is proving difficult to advance, as there is no agreed baseline understanding about the subject in general or detailed definitions in particular: LAWS means different things to different people. This difference in understanding may range from a fanciful Hollywood-generated image of a post-apocalyptic future to a fear that some current weapon systems may be more than they seem. Previous decades abound with learned articles predicting a future dominated by machines, to the detriment of mankind; the ICRC was predicting such a future over 30 years ago. However, what we actually learn is that few, if any, of these doomsday scenarios have come to pass. What we do know, however, is that warfare remains a fundamentally human activity, although the way in which it is conducted has changed a great deal over the years. We have seen developments in automation within systems, and indeed the development of unmanned systems operated from distant locations. What we have seen has been an evolving intelligent partnership between operators and computers, not a paradigm shift.

Without some hard definitions to anchor the debate surrounding LAWS, it will continue to drift between polarized views, effectively going nowhere. The UK, in a Joint Doctrine Note published in 2011 (currently being updated) offers the following definitions of two concepts central to the debate, automation and autonomy:

- An **automated or automatic system** is programmed to logically follow a predefined set of rules with predictable outcomes, e.g. the Phalanx anti-ship missile defence system, or a remotely piloted aircraft system programmed to return to a fixed point after a signal outage.
- An **autonomous system** is capable of understanding higher-level intent and direction. From this understanding, as well as a sophisticated perception of its environment, such a system is able to take appropriate action to bring about a desired state. It is capable of deciding on a course of action from among a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous system will be predictable, individual actions may not be (i.e. the system does not merely follow a pattern of rules in a predictable way).

We argue that, while weapons will become increasingly automated, it is extremely unlikely that technology – including all the work currently underway in artificial intelligence – will be able to produce a system capable of understanding intent, judged against a sophisticated perception and understanding of the environment in which it is operating.

Furthermore, not only do we doubt that a machine will be capable of this sophisticated level of understanding, it is very unlikely that a machine would be able to operate in the demanding and high-pressure environment of military decision-making. Military decision-making is guided by professional judgement gained from experience, knowledge, education, intelligence and intuition. Commanders continuously combine intuitive and analytic approaches to decision-making. The intuitive approach is the act of reaching a conclusion that emphasizes pattern recognition based on knowledge, judgement, experience, education,

intelligence, boldness, perception and character. The analytic approach aims to produce the optimal solution to a problem from among the solutions identified. The intuitive and analytic decision-making approaches are not, therefore, mutually exclusive. Commanders may make an intuitive decision based on a situational understanding gained during the planning. If time permits, the staff may use specific tools, such as war gaming, to validate or refine the commander's intuitive decision.

Decision-making in a high-stakes, high-pressure situation is never easy. It can make or break an organization and its leadership. This level of understanding and learning is beyond the learning predicted for machines, where known and new variables are judged against a set of finite numbers of possible outcomes. In the military environment, commanders face a thinking and adaptive enemy. Commanders estimate, but cannot predict, the enemy's actions and the course of future events.

To make effective decisions, there must be a clear understanding of personal and organizational values. When ideals are shared and mutually understood, this can help people to focus on the most important issues and to make decisions. Integrating values into more routine decisions helps teams internalize them before a crisis hits. It is common to look at the long-term implications of options. However, it is also critical to regularly consider the impact of each choice on the perception of an organization's credibility and other core values. As with many armed forces, the UK recognizes that we now live in a world where winning the "battle of the narratives" is as important as winning the tactical battle. The narrative is often based on our core values and beliefs, which in the UK's case are closely allied to a strong desire to abide by and uphold the rules of international humanitarian law (IHL). Therefore, would any commander place his or her own reputation, and potentially the nation's, in the hands of a machine whose actions are neither predictable nor accountable?

Often key military decisions, especially those involving kinetic action, are judged against several very human attributes:

- the need to understand the downside of any decision, that is, the level of possible regret;
- the need to understand how, if and when a decision may be repealed;
- the need to understand the possible repercussions of a decision. This includes second- and third-order effects; and
- the effect of a decision on personal and organizational resilience, and what impact that may have on subsequent decision-making and flexibility.

It is unlikely that a machine will ever be capable of understanding and acting on these and many other human factors.

With regard to the phrase "meaningful human control", it is one that the UK does not currently use in either our policy or our doctrine; however, as the phrase has now gained some currency in the LAWS debate, it is worth reinforcing two major principles of the UK's operational doctrine, which emphasizes the level of human input into the targeting process:

- The decision on whether to use lethal force against a legitimate military target is made through a rigorous human-led targeting process, the targeting directive for a specific theatre of operations and rules of engagement.
- In practical terms, every target is assessed by a human, and every release of weapons is authorized by a human. All weapon use is, and will continue to be, based on rules of engagement which seek to ensure adherence to IHL.

The UK is of the opinion that the way to regulate all current and future weapon systems is through a rigorous Article 36 review, conducted by lawyers, as required by Protocol I

additional to the Geneva Conventions. This instrument obliges the High Contracting Parties, in the study, development, acquisition or adoption of a new weapon, means or method of warfare, to determine whether its employment would, in some or all circumstances, be prohibited by Additional Protocol I or by any other rule of international law applicable to the High Contracting Parties. Specifically, such a review should determine whether a weapon:

- is prohibited, or whether its use is restricted, by any specific treaty provision or other applicable rule of international law;
- is of a nature to cause superfluous injury or unnecessary suffering;
- is capable of being used discriminately;
- is expected to cause widespread, long-term and severe damage to the natural environment; and
- is likely to be affected by current and possible future trends in the development of IHL.

Currently, only some 25 States are known to conduct formal Article 36 reviews; conceivably, if more States were to adopt the practice, the fear, suspicion and misunderstanding surrounding the development of future weapon systems could be partly allayed. Article 36 reviews have been capable of dealing with advances in technology for close to 40 years; there is no reason to doubt their suitability for dealing with greater advances in autonomy.

Stepping outside the recognized regulatory mechanism of the Article 36 review and looking at additional regulation, which in the case of LAWS would almost certainly be pre-emptive, creates a number of dangers. Much of the research concentrating on increased precision and target discrimination to reduce collateral damage and improve effectiveness could be affected by such a wider ban. The research on and development of highly automated defensive systems could also suffer, and research into autonomy in non-lethal areas, such as logistics or information management, could be similarly stifled. Furthermore, any pre-emptive regulations may spill over into the civil sector, disrupting the development of and research into a wide range of civilian applications in advanced technologies.

The UK remains one of a very small number of States to publically record its policy concerning LAWS. This policy makes it clear that the operation of our weapons will always be under human control as an absolute guarantee of human oversight, authority and accountability for weapons usage. Moreover, the UK Government does not possess fully autonomous weapon systems, and has no intention of developing them. While a limited number of defensive systems can operate in highly automated mode, there is always a person involved in setting the parameters of any such mode.

In the operational environments of the future, which are likely to be characterized as contested, congested, cluttered, connected and constrained, the need to ensure that humans retain control of military decision-making and weapon use is unlikely to diminish, especially when set against the need to win the battle of the narratives.

Russia's automated and autonomous weapons and their consideration from a policy standpoint

Speaker's summary

Dr Vadim Kozyulin, PIR Center for Policy Studies, Russian Federation

Russia has a long history of producing automated weapons for air defence, the air force, army and navy and the Russian strategic missile forces. Ground-to-ground, air-to-air, anti-warship and other types of missiles, active-protection weapons and torpedoes with a high degree of automation are not regarded as autonomous weapon systems, but as automatic weapons with various guidance systems, as well as engagement and activation programmes. There is no definition of autonomous weapon systems in the military encyclopaedia that can be found on the website of the Russian Ministry of Defence. Instead, one can find the term "combat robot".

"A combat robot is a multifunctional device **with anthropomorphic (humanlike) behaviour** that partially or fully **performs the functions of a human** during particular combat missions."[17] I assume that the words "anthropomorphic (humanlike) behaviour" are the key to understanding what the Russian military means when it talks about autonomous weapon systems.

The military encyclopedia divides robots into three generations:

- first-generation robots with software and remote control can only function in an organized environment;
- second-generation robots are adaptive, having some kind of sensory organ and being able to operate in previously unknown conditions, i.e. to adapt to a changing environment; and
- third-generation robots are intelligent, having a control system with elements of artificial intelligence (existing only in the form of laboratory models for now).

Based on this classification, we may assume that third-generation robots might be referred to as autonomous weapon systems from the Russian military's point of view.

In reviewing various international publications, some observers might conclude that Russia has advanced very far in the development of autonomous weapon systems. In May 2014, *Popular Mechanics* wrote: "While research stalls in the United States, Russia's leaders are determined to make their country a robot superpower." This point of view is not shared by the Moscow Technical University bulletin, which wrote in 2013: "The gap between Russia and the USA in the field of automation of military hardware currently amounts to about 10–15 years. The main problem, which largely determines this lag, is the lack of a developed State technical policy in this sphere."

For a balanced judgement, we may consider weapons that are at the edge of autonomous weapon system technologies, i.e. unmanned aerial vehicles or drones. The United States made the Predator in 1994, and the first missile launches by a Predator took place in 2001. Russia has no armed drones for now, although they are being developed, which shows that Russian drone designers are about 15 years behind their American counterparts.

[17] See http://encyclopedia.mil.ru/encyclopedia/dictionary/details.htm?id=3551@morfDictionary.

Now I'll say some words about the efforts that Russia has made to catch up with the global trend towards developing autonomous weapon systems. The Robotization of Weapons and Military Equipment 2015 programme was adopted in 2000. Specialists estimate that it allowed for the design of certain experimental samples and equipment prototypes. However, further work and research stopped because of financial constraints, and the Russian defence design bureaus needed a new campaign that could breathe life into their ageing structures.

In September 2015, the Russian Defence Ministry developed the Comprehensive Policy "Programme for Development of Advanced Military Robotics up to 2025 with Forecasts until 2030", reflecting the main thrusts in the development of robotic systems for military purposes for all branches of the Russian armed forces. Representatives of the Defence Ministry regard the robotization of arms, military and special equipment as priority areas for the development of the armed forces, including the development of unmanned vehicles in the form of robotic systems and complexes for the military to be applied in various environments.

In 2015, the Russian General Staff adopted the "Concept for Deployment of Robotic Systems for Military Use until 2030". The document sets out general technical guidelines for ground-based robotic systems for military purposes. The State standards establishing uniform requirements for military robotics are currently under development in four main areas:

1. unmanned aerial vehicles
2. ground-based robotic systems
3. underwater autonomous vehicles
4. unmanned boats.

Both documents are classified.

On 16 December 2015, President Putin signed the Decree "On the National Centre for Technology Development and the Basic Elements of Robotics", establishing the main institution for the design of Russian autonomous weapon systems. I would like to point out that these fundamental steps for the Russian Defence industry were taken at the end of 2015, which clearly demonstrates that Russia is rather at the beginning of the road.

To illustrate this, I will share with you some facts about the most modern Russian autonomous weapon system deployed by the Russian armed forces. At the moment, the Russian military use unmanned aerial vehicles and two types of ground-automated systems. The number of unmanned aerial vehicles increased from 180 in 2011 to 1,780 in 2016.

The most popular unmanned aerial vehicles are the Orlan-10 and the Eleron-3SV (T23E UAV). The two ground-automated systems are the Raznoboy mobile robotic complex for radiation reconnaissance and transportation of radioactive materials, and the Berloga-P remote-controlled radiation and chemical reconnaissance vehicle.

The Orlan-10 is an unmanned surveillance and reconnaissance aircraft that enables objects on land to be monitored in search, reconnaissance and rescue operations. The aircraft is resistant to adverse weather conditions and has a modular architecture that allows for quick changes in the composition of onboard equipment. The Orlan-10 is able to carry both photo and video cameras, and a thermal camera or signal repeater.

With a maximum take-off weight of 3.5 kg, the Eleron-3SV is capable of carrying special TV, photo or infrared equipment, a repeater, electronic intelligence stations and jamming equipment, delivering it to the target area and returning to a landing site in accordance with the predetermined route. The unmanned aerial vehicle returns to the landing site automatically. It can perform visual inspection of combat training, monitor the state of military infrastructure, etc. The range is about 25 km with a flight duration of 90–120 minutes, and the

flight altitude is up to 3 km. The Eleron-3 can fly in stand-alone mode as well as by radio command.

Two ground vehicles deployed by the Russian army are:

- The Berloga-P remote-controlled radiation and chemical reconnaissance vehicle. Deployed by the Russian army in 2004, it is designed for radiation and chemical reconnaissance, including searching for gamma-radiation sources in hard-to-reach areas in industrial and residential buildings. The system includes a land vehicle, equipped with a manipulator, a television system, radiation and chemical reconnaissance equipment and remote control via a radio command system, as well as a data-collection and information-processing unit.

- The Raznoboy mobile robotic complex for radiation reconnaissance and transportation of radioactive materials was deployed by the Russian army in 2003. It is designed for visual and radiation reconnaissance, gamma search, sampling and transportation of solid radioactive materials during operations in highly contaminated areas by post-accident clean-up groups and units. The complex consists of a four-wheel-drive truck-type KAMAZ 43114 trailer, two mobile modules (MRK-46M and MRK-PX), a remote control centre, a communications unit and various auxiliary equipment (soil and liquid samplers, a demolition hammer, a rotary hammer, fork trucks, etc).

There are no unmanned vessels deployed by the Russian navy.

One other autonomous weapon system is the Cobra-1600 Light Sapper Robot, which is expected to be deployed in 2016. The Cobra-1600 is designed to conduct remote visual exploration, search and primary diagnosis of potential explosive devices using a TV camera, special equipment and remote-defusing improvised explosive devices (IEDs), and to load IEDs into special containers for evacuation, as well as for operations providing access to potentially dangerous objects.

Now I'll give you some examples of autonomous weapon systems being tested by the Russian Defence Ministry. First are two vehicles from the Uran family:

- The Uran-6 is an unmanned multifunctional minesweeping system powerful enough to replace 20 professional sappers, and it can be remotely operated from a safe distance of about 1.5 km. The robot system is capable of distinguishing an air bomb from an artillery shell or an anti-tank mine. The Uran-6 is still not fully reliable, and testing is followed by sappers who verify how efficiently it has cleared the zone.

- The Uran-14 MRTK-P robotic mine-clearance and firefighting tracked vehicle is designed for minesweeping and for extinguishing fires in life-threatening, hazardous environments and inaccessible areas. A sophisticated video system allows the operator to have full control of the vehicle movement during its operation. The video system consists of six high-resolution and waterproof cameras. One thermal camera allows the Uran-14 to operate in reduced visibility conditions. A Global Positioning System-Global Navigation Satellite System and inertial navigation system permits the Uran-14 to return to the initial location should radio communication be lost.

Now I will discuss combat platforms that are being tested by the Russian army.

In 2015, Rostec unveiled the Uran-9 unmanned combat ground vehicle. The system will be designed to deliver combined combat, reconnaissance and counter-terrorism operations as well as fire support. One unit consists of two reconnaissance and fire-support robots, a

tractor for their transportation and a mobile control post. The armament of the reconnaissance and fire-support robots may vary depending on customer requirements. The robots are fitted with a laser-warning system and target-detection, identification and tracking equipment.

The multifunctional robotic system Nerehta has a modular design whereby equipment or weapons can be changed depending on the mission. It is designed so that it can also be used for evacuating the wounded, patrolling and guarding strategic objects, extinguishing fires, remote mining, and cutting wire and breaching obstacles. I assume that the engineers are currently considering the vehicle's application, and testing should start in 2016.

Platforma-M is a remote-controlled robotic unit on a crawler, armed with grenade launchers and Kalashnikov rifles. It is equipped with optical-electronic and radio-reconnaissance locators, which enable the robot to perform combat tasks at night. Platforma-M is a universal platform that can be supplied with a variety of chassis and weaponry. The electric battery is sufficient for 48 hours of operation. The robot is being tested by Russian marines against simulated intruders during anti-terrorist drills.

Conclusions

1. The Russian Army has only remote-controlled robots at the moment.
2. They can be partially autonomous.
3. They are reprogrammable, which allows for a higher degree of autonomy in the future.

The Russian Defence Ministry considers that combat robots can perform the functions of humans, inasmuch as:

- they are capable of replacing people in hostile environments, such as conditions of high radiation, chemical contamination, extremely high or low temperatures, etc.;
- they can assist Russian frontier guards in protecting national borders (60,933 km long, including 38,808 km of sea borders);
- robots can save the lives of solders in minefields, in fires and under fire; and
- the decision to use lethal force is ultimately under human control.

The Russian Defence Ministry pays special attention to the 1949 Geneva Conventions and Additional Protocol I. The provisions of these documents are reflected in general terms in the internal service regulations of the armed forces, and in detail in the "Instruction for Application of Humanitarian Law in the Armed Forces", drafted by the Army Department for Combat Training. Russian military schools have a special 80-hour course in their curriculum devoted to international humanitarian law. Officers in Russian military academies have to attend a similar 60-hour course repeatedly. ICRC representatives regularly deliver lectures on humanitarian law to the Russian military and occasionally take part in manoeuvres as observers.

The issue of autonomous weapon systems is rather new to the Russian expert community, although it is gradually gaining public attention. Russian military analysts discuss the following issues in their publications and at relevant meetings:

1. terminology in the field of autonomous weapon systems;
2. guidelines for the practical use of ground autonomous weapon systems;
3. setting up common standards for autonomous-weapon-system control software, information exchange and processing;
4. responsibility for atrocities committed by autonomous weapon systems;

5. limits and rules for future artificial intelligence in order to prevent its use in ways inimical to mankind; and
6. security measures to prevent autonomous weapon systems from falling into the hands of terrorists.

My perception of the potential threats posed by autonomous weapon systems is as follows:

1. One day such weapon systems might develop to a level that could pose a threat of occasional incidents or accidents in the air, at sea or in deep waters. Preventing incidents caused by these systems might therefore become an issue to address.
2. Incidents might happen because of loss of communication, jamming or control interception – in other words, because of communications-security or cyber-security failures.
3. There are signs that the expansion of combat autonomous weapon systems could lead to a new arms race.
4. At the end of the nineteenth century, military commanders in many European countries were impressed by the new opportunities afforded by two modern technologies: railways and mobilization plans. Railways and mobilization plans allowed them to quickly deliver masses of solders to the front line. Although this happened accidentally, once it started, the process could not be reversed, and this is how the Second World War actually began.

There is no doubt that combat robots will significantly increase the military potential of military powers, and the mass adoption of lethal autonomous weapon systems could pose a threat to the military balance in different regions, thereby increasing the risks of conflict. It is crucial for mankind to assess and neutralize this threat in order to allow artificial intelligence and autonomous robots to be synonymous with a better and safer world, and not with an unknown and irreversible danger.

Addressing the challenges raised by increased autonomy

Speaker's summary

Ms Kerstin Vignard, United Nations Institute for Disarmament Research

These last two years of discussions on lethal autonomous weapon systems have been rich and educational. We have had the opportunity to interact with experts on the technical, legal, ethical and other aspects of the issue. These rich exchanges have been productive in helping the international community to get its bearings on this issue. The upcoming week of informal discussions is particularly important, as it will help to delineate the area of future debate as we move towards the Review Conference of the States Parties to the Convention on Certain Conventional Weapons (CCW).

So, in the spirit of encouraging governments to move the conversation forward, I would like to offer for consideration five observations, the first and last of which are about alternative frames for this issue.

1. "Autonomous weapon system" is a misnomer

In 2014, the United Nations Institute for Disarmament Research (UNIDIR) published its first observation paper, entitled "Framing Discussions on the Weaponization of Increasingly Autonomous Weapon Systems".[18] Framing discussions in a productive way can help focus them on the critical issues and set them up for success.

Thus, the first observation is that it is time to take the international discussion on lethal autonomous weapon systems and, at a minimum, reframe it as "autonomy IN weapon systems". This isn't just playing with words. It acknowledges that varying levels of autonomy might be applied to different characteristics within the same object or weapon system. This frame also allows us, in the short term, to pivot away from trying to draw – and agree upon – lines between fully versus semi-autonomous weapons or those with supervised autonomy, or from attempts to come up with a formula for determining whether something is autonomous or just highly automatic. Instead, it allows us to focus on the functions that, when increasing amounts of autonomy are applied to them, raise concerns and challenges, or show us where we need to develop shared understandings of how our existing obligations and norms apply.

2. The whole will be greater than the sum of its parts

Many States affirm that autonomy discussions within the framework of the CCW are not about existing systems. However, there are highly automatic or autonomous components or features of existing systems, which, if combined in particular ways in the future, may pose new and unique concerns, even if the features or weapons themselves are not problematic today. This moves the timeline from a far-off future concern to a much more near-term technological possibility.

A dynamic exchange between States on how these existing or near-term features might combine, and the acceptability of these different combinations, would be useful. Some are perhaps uncomfortable about discussing existing systems. However, this discussion is

[18] UNIDIR, *Framing Discussions on the Weaponization of Increasingly Autonomous Weapon Systems*, UNIDIR Resources No. 1, 2014, http://www.unidir.org/files/publications/pdfs/framing-discussions-on-the-weaponization-of-increasingly-autonomous-technologies-en-606.pdf.

different than, and shouldn't be confused with, the categories that States might decide to eventually regulate or control. Understanding and affirming the particular areas that States do not see as problematic will help to clear space and enable us to focus on those areas that might be problematic, or those where States might have some uncertainty. Starting from existing, widely accepted, highly automatic systems, as we "dial up" autonomy for different parameters (mobility, time of autonomous operation, etc.[19]), at which point do specific legal, technical, operational or ethical concerns arise? Mapping these friction points in a more systematic way will bring much more focus to the international discussions. We acknowledge that this would be a complex discussion. However, in the absence of a definition, it would also allow the conversation to be narrowed down to potentially problematic applications of autonomy.

There is a second area where the whole is greater than the sum of its parts: increasingly autonomous systems working in concert with other increasingly autonomous systems. We must not lose sight of how features that are connected, interactive and increasingly autonomous might further attenuate human control or intent.

3. We see a real gap in many policymakers' understanding in two relevant areas: learning systems and interactions with cyber operations

A much deeper understanding of learning systems – their capabilities and their limitations – is necessary to ensure that we are developing sound policy. In this regard, the concept of predictability in learning systems needs much greater attention. We need to unpack what is meant by "predictability": in the goal? In the sub-goals? In the means of achieving goals and sub-goals? More discussion on the challenges that learning systems would pose to weapon reviews is also needed.

We will also eventually need to consider that autonomous weapons are not limited to conventional weapons. At some point, States will confront the issue of whether the concerns we have about the weaponization of increasingly autonomous technologies are also applicable to increasingly autonomous intangible technologies, such as cyber operations, particularly if these can have kinetic effects.

4. Check your blind spots – keep alternative developmental trajectories in sight

There are two ways to develop AWS:

- building from scratch or adapting a remotely controlled or automatic system and empowering it to select and attack targets autonomously (for example, through more capable sensors, programming, or processing power); or
- weaponizing a civilian autonomous technology.

Most States interested in increasing autonomy are likely to use the first path. And discussions within the framework of the CCW are likely to focus predominantly on preparing to address this developmental trajectory. However, it is the second trajectory that will be more challenging to control and respond to, because it is more likely to take us by surprise.

Thus, we could usefully pay more attention to increasingly autonomous technologies in the civilian sector. In contrast to a gradual development of increasing autonomy under military control and oversight, it is possible that a party may take a completely civilian technology and weaponize it.

[19] UNIDIR, *op cit.*, pp. 4–5.

This developmental trajectory is not limited to conventional forces; non-State entities, terrorists, extremists, criminals and individuals may also follow it. This trajectory is characterized by:

- creativity. Civilian autonomous devices could be "weaponized" and used in novel ways that are "unthinkable" until they happen (prior to 9/11, it was mostly novelists, conspiracy theorists and screenwriters who imagined how commercial aircraft could be used as projectiles in a mass casualty attack);
- being technologically accessible (we are talking about off-the-shelf, relatively low-cost civilian technology, for which access to components or open-source information is difficult to control); and
- having lower standards of reliability and predictability, thus posing a much graver threat to international humanitarian law (IHL) and human rights, as they are likely to be much cruder weapons.

These challenges are, of course, not unique to autonomous weapons. We struggle to respond effectively to improvised explosive devices for many of these same reasons.

So, while this is not the immediate focus of the discussions under the CCW, one of the most important things we can do is to start consolidating norms on the weaponization of increasingly autonomous technologies. Norm development is already starting outside the CCW – consider the open letter on artificial intelligence coordinated by the Future of Life Institute. This is another reason to step up the pace and focus of international discussions.

5. Putting the horse before the cart – or, in this case, the human before the robot

We all agree that there are two sides to the autonomy issue – the technology side and the human side. In discussions on autonomous weapon systems, the tendency is to start with discussion of the technology.

Technology moves quickly, and humans move slowly. Any approach that attempts to predict or regulate the evolution of the rapidly moving fields of technology, computation, robotics and artificial intelligence will always remain – pardon the pun – a moving target.

To put it another way, the conversation has been set up in a reactive way, reacting to technology, whether existing, potential or imagined. We suggest that the technology should not be guiding the conversation. We need to shift to a proactive conversation on human control, judgement, intent, responsibility, etc., and then apply that to technology.

Ultimately, the autonomy question is really about what control or oversight we expect humans to maintain over the tools of violence that we employ. In UNIDIR's first observation report, we put it as follows:

> Rather than trying to agree upon rigid categories or definitions of thresholds of autonomy, in the initial stage of discussions, States might consider focusing discussion on identifying the critical functions of concern and the interactions of different variables. This would anchor the discussion and set its boundaries. It would also allow discussions to bypass—for the time being—getting bogged down into a technology-centric definitional exercise.[20]

[20] UNIDIR, *op cit.*, p. 5.

Later that year, in our second observation report,[21] we noted that focusing on the human side of the autonomy question, rather than a technological framing of the issue, has several benefits:

- it provides a common language for discussion that is accessible to a broad range of governments and publics regardless of their degree of technical knowledge;
- it focuses on the shared objective of maintaining some form of control over all weapon systems;
- it is consistent with IHL regulating the use of weapons in armed conflict, which implicitly entails a certain level of human judgement and explicitly assigns responsibility for decisions made; and
- it is a concept broad enough to integrate consideration of ethics, human-machine interaction and the "dictates of the public conscience", which are often sidelined in approaches that narrowly consider just technology or just law.

Reaffirming the principles of human control and judgement, and getting down to work on developing a shared understanding of how this applies specifically to the weaponization of increasingly autonomous technologies (i.e. how and when human control and judgement are exercised and what makes this meaningful or appropriate), are the urgent next steps for the international community. Of course it is not possible to have the "human" conversation in a vacuum – we will need to talk about technologies. However, we can use specific technologies to illustrate or test these human principles and ensure that our values and principles continue to be embedded in decisions about weapon development and design, and about their eventual use.

As an additional benefit, this approach will be applicable to as-yet unimagined technological developments that we might consider weaponizing in the future. At a time when scientific understanding and technology are developing at exponential rates, in surprising and non-linear ways, these shared principles will be increasingly necessary.

In conclusion, when approaching the autonomous-weapon-system issue within a tech-centric frame, there is a temptation to imagine these weapon systems as something other than simply another tool for us to select in order to achieve specific strategic and operational objectives. We must resist this temptation, as it clouds both the discussion and perhaps even our judgement. Automated weapon systems will not be our peers, as they are not human. They will not be our fellow soldiers, no matter how well integrated they are in our military units. They are our tools.

The importance of your work under the CCW cannot be overstated. It will ensure that we do not cede – even unintentionally – our legal or moral responsibilities, nor our humanity, to an object, no matter how technologically sophisticated or capable it is. Putting the human side of the equation first, rather than the technology side, helps us to keep this fundamental distinction at the forefront of our discussions.

[21] UNIDIR, *The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward*, UNIDIR Resources No. 2, 2014, p. 3, http://www.unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf.

PART III: BACKGROUND PAPER PREPARED BY THE INTERNATIONAL COMMITTEE OF THE RED CROSS, MARCH 2016[1]

CONTENTS

[1] **This is an edited version of the background paper circulated to participants in advance of the ICRC's expert meeting**. It was drafted by Dr Neil Davison, Scientific and Policy Adviser, and Dr Gilles Giacca, Legal Adviser from the Arms Unit in the Legal Division at the ICRC.

1. INTRODUCTION

Debates on autonomous weapon systems have expanded significantly in recent years in diplomatic, military, scientific, academic and public forums. These have included expert discussions within the framework of the UN Convention on Certain Conventional Weapons (CCW) in 2014, 2015 and 2016, and an expert meeting convened by the International Committee of the Red Cross (ICRC) in 2014.[2]

Views on this complex subject, including those of the ICRC, continue to evolve as a better understanding is gained of current and potential technological capabilities, the military purpose of autonomy in weapon systems, and the resulting questions for compliance with international humanitarian law (IHL) and ethical acceptability.

Discussions among government experts have indicated broad agreement that "meaningful", "appropriate" or "effective" human control over weapon systems must be retained, for legal, ethical and/or policy reasons. The ICRC, for its part, has called on States to set limits on autonomy in weapon systems to ensure that they can be used in accordance with IHL and within the bounds of what is acceptable under the dictates of public conscience.[3]

In view of the incremental increase of autonomy in weapon systems, specifically in the "critical functions" of selecting and attacking targets, experience with existing weapon systems can provide insights on where the limits on autonomy in weapon systems should lie, and the kind and degree of human control that may be deemed meaningful, appropriate or effective.

As a further contribution to this discussion, the ICRC convened this second international expert meeting, entitled "Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons", in order to:

- discuss the characteristics of autonomous weapon systems;
- understand autonomy in the critical functions of existing weapon systems;
- explore emerging technology and the implications for future autonomous weapon systems;
- examine the legal and ethical implications of increasing autonomy in weapon systems;
- consider the legal, military (operational) and ethical requirements for human control over weapon systems and the use of force; and
- share possible approaches to addressing the challenges raised by increasing autonomy.

This background paper was provided in order to guide participants on the key areas for discussion at the expert meeting. Except where explicitly mentioned, the paper does not necessarily represent institutional positions of the ICRC. References to specific weapon systems are based on limited publicly available sources and are provided for illustrative purposes only.

---

[2] ICRC (2014) *Autonomous weapon systems technical, military, legal and humanitarian aspects*, Report of an Expert Meeting held 26-28 March 2014 (published November 2014), https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf.

[3] ICRC (2015) *Statement to the 2015 Meeting of High Contracting Parties to the Convention on Certain Conventional Weapons*, 13 November 2015, http://www.unog.ch/80256EDD006B8954/(httpAssets)/3E46D595853F5361C1257F0F004C22C3/$file/ICRC+CCW+AWS+statement+FINAL.pdf.

2. CHARACTERISTICS OF AUTONOMOUS WEAPON SYSTEMS

There is no internationally agreed definition of an autonomous weapon system, but common to various proposed definitions is the notion of a weapon system that can independently select and attack targets.[4]

The ICRC has suggested that "**autonomous weapon system**" is an umbrella term that encompasses any weapon system, wherever it operates, 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. After initial activation by a human operator, it is the weapon system itself – using its sensors, computer programming and weapon(s) – that takes on the targeting processes and functions that are ordinarily controlled directly by humans.

The ICRC's working definition includes all weapons in which these critical functions are performed by the sensors and programming of the weapon system, rather than directly by a human operator. At a fundamental level, it is autonomy in the critical functions that distinguishes autonomous weapon systems from all other weapons, including those in which these functions are remotely controlled by a human operator.

Some States and other experts have made a distinction between "highly automated weapon systems" and "fully autonomous weapon systems" based on the degree of freedom of the weapon to determine its own functions, and with "fully autonomous" assuming the machine can set its own goals, or even "learn" and adapt its functioning. However, the ICRC's working definition, which it submitted to frame the discussion at this expert meeting, encompasses any weapon that can independently select and attack targets, whether described as "highly automated" or "fully autonomous".

The rationale for the ICRC's approach is that all weapons with autonomy in the critical functions raise the same core legal and ethical questions:

- In the intended circumstances of use, can the weapon system select and attack targets in a way that respects the **rules of IHL**?
- In cases where operation of the weapon system results in an apparent violation of IHL, is it possible to attribute **responsibility** to an individual or a State, and to hold them **accountable**?
- Is it **ethically acceptable** (based on the principles of humanity and the dictates of the public conscience) for the weapon system to independently select and attack targets?

To summarize, for the purposes of the expert meeting, the **working definition** of an autonomous weapon system is:

Any 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.[5]

[4] ICRC (2014) *Autonomous weapon systems technical, military, legal and humanitarian aspects*, (footnote 2), pp 63-64.
[5] ICRC (2015) *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 (published October 2015), pp 44-47, https://www.icrc.org/en/download/file/15061/32ic-report-on-ihl-and-challenges-of-armed-conflicts.pdf.

3. AUTONOMY IN EXISTING WEAPON SYSTEMS

Examples of existing weapon systems with some autonomy in their critical functions include: missile- and rocket-defence weapons; vehicle "active-protection" weapons; anti-personnel "sentry" weapons; sensor-fused munitions, missiles and loitering munitions; and torpedoes and encapsulated torpedo mines.

These weapons vary greatly in their mechanisms of operation and operational uses. However, many are highly constrained in the tasks they are used for (e.g. defensive rather than offensive operations), the types of targets they attack (vehicles and other objects rather than personnel), the circumstances in which they are used (in simple, relatively predictable and constrained environments rather than complex, unpredictable environments), and/or the time frame of their autonomous operation. Some are also supervised in real time by a human operator.

3.1 Missile- and rocket-defence weapons

These weapons are used for air defence, including short-range defence of ships or ground installations against missiles, rockets, artillery, mortars, aircraft, unmanned systems and high-speed boats. Ship-based weapons are often described as "close-in weapon systems" and the land-based weapons as "counter-rocket, artillery and mortar systems". Once activated, these weapons carry out target selection and attack without further human intervention, but remain under the overall supervision of human operators who may be able to deactivate them following a malfunction or an unintended attack. Some States categorize them as "human-supervised autonomous weapon systems",[6] and others have described them as highly automated weapon systems.

Generally, these weapons incorporate a radar, to detect incoming projectiles or aircraft, and a computer-controlled "fire-control system" to aim and fire the weapon. The weapon system selects an incoming projectile, estimates its trajectory, and then fires interceptor missiles or bullets (depending on the system) to destroy it. Some of these weapons have been in use since the 1980s and are in the inventories of at least 30 States.[7] Illustrative examples include:

- ***Goalkeeper Close-in Weapon System*** *(Netherlands),* a ship-based 30 mm gun system;[8]
- ***Iron Dome*** *(Israel)*, a mobile land-based counter-rocket, artillery and mortar system that fires interceptor missiles;[9]
- ***Kashtan Close-in Weapon System*** *(Russia)*, a ship-based combined 30 mm gun and missile system;[10]
- ***Nächstbereichschutzsystem (NBS) MANTIS*** *(Germany)*, a land-based 35 mm counter-rocket, artillery and mortar system;[11]

[6] For example, US Department of Defense (2012) *Autonomy in Weapon Systems, Directive 3000.09,* 21 November 2012.
[7] P Scharre and M Horowitz (2015) *An Introduction to Autonomy in Weapon Systems*, Center for a New American Security, February 2015.
[8] Thales, *Goalkeeper Close-in Weapons System*. According to the manufacturer it is deployed by Belgium, Chile, the Netherlands, Portugal, Qatar, South Korea, the UAE and the UK, https://www.thalesgroup.com/en/netherlands/defence/goalkeeper-close-weapon-system.
[9] Rafael Advanced Defense Systems, *Iron Dome Dual-Mission Counter Rocket, Artillery and Mortar (C-RAM) and Very Short Range Air Defense (V-SHORAD) System*, http://www.rafael.co.il/Marketing/186-1530-en/Marketing.aspx.
[10] Navy Recognition, *Kashtan, Kashtan-M, CADS-N-1, Palma, Palash close in weapon systems (CIWS)*, http://www.navyrecognition.com/index.php/east-european-navies-vessels-ships-equipment/russian-navy-vessels-ships-equipment/weapons-a-systems/123-kashtan-kashtan-m-kashtan-lr-cads-n-1-close-in-weapon-system-ciws-.html.
[11] Army Technology*, NBS Mantis*, http://www.army-technology.com/projects/mantis/

- ***Phalanx Close-in Weapon System*** *(USA)*, a ship-based 20 mm gun system.[12] Similar technology is used for the **Counter-Rocket, Artillery, and Mortar System (C-RAM)**, a land-based version of the system, and the **SeaRAM Close-in Weapon System**, a ship-based system that fires interceptor missiles.[13]
- ***Type 730*** *and* ***Type 1130 Close-in Weapon System*** *(China)*, ship-based 30 mm gun systems.[14]

3.2 Vehicle "active-protection" weapons

Vehicle "active-protection" weapons are designed to protect armoured vehicles from attacks with missiles, rockets, rocket-propelled grenades and other projectiles.[15] Once activated, the weapon system selects and attacks the incoming projectile without human intervention. Some of these systems appear to operate in a similar way to counter-rocket, artillery and mortar systems (Section 3.1). Illustrative examples include:

- ***Advanced Modular Armour Protection Active Defence System (AMAP-ADS)*** *(Germany)* selects incoming projectiles and attacks them with a directed blast;[16]
- ***Arena*** *(Russia)* uses a radar to select incoming projectiles and fires rockets to destroy them;[17]
- ***LEDS-150*** *(South Africa)* detects laser rangefinders associated with anti-tank weapons and can be configured to fire autonomously to intercept an incoming projectile;[18]
- ***Quick Kill*** *(USA)* uses a radar to select incoming projectiles and fires an interceptor missile to destroy them;[19]
- ***Trophy*** *(Israel)* uses a radar to select incoming projectiles and fires a directed blast to destroy them;[20]
- ***Zaslon*** *(Ukraine)* uses a radar to select incoming projectiles and fires munitions to destroy them.[21]

3.3 Anti-personnel "sentry" weapons

Anti-personnel "sentry" weapons, for use at specific sites, perimeters or borders, have been developed to have increasing levels of autonomy in the critical functions of selecting and attacking targets. Some systems have been deployed and others are still in development. Some are stationary defensive systems and others are mobile systems that might patrol an area. Existing "sentry" weapon systems appear to select targets autonomously, but require remote authorization from a human operator to attack. However, some reports have

[12] Raytheon, *Phalanx Close-In Weapon System*. According to the manufacturer it is deployed by 25 States, http://www.raytheon.co.uk/capabilities/products/phalanx/;
US Navy (2013) *MK 15 - Phalanx Close-In Weapons System (CIWS)*, 15 November 2013, http://www.navy.mil/navydata/fact_print.asp?cid=2100&tid=487&ct=2&page=1.
[13] US Navy (2013) *SeaRAM Close-In Weapon System (CIWS) Anti-Ship Missile Defense System*, 15 November 2013, http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=456&ct=2.
[14] J Bussert (2013) China's Navy Deploys Three-Tier Defensive Weapons, *Signal Magazine*, http://www.afcea.org/content/?q=china%E2%80%99s-navy-deploys%E2%80%A8-three-tier-defensive-weapons.
[15] The Economist (2011) The armour strikes back, *The Economist*, Technology Quarterly Q2, 2 June 2011, http://www.economist.com/node/18750636.
[16] Rheinmetall Defence, *Protection Systems Land*, http://www.rheinmetall-defence.com/en/rheinmetall_defence/systems_and_products/protection_systems/protection_systems_land/index.php.
[17] Konstruktorskoye byuro mashynostroyeniya, *Arena Active Protection System*, http://www.kbm.ru/en/production/saz/368.html.
[18] Saab, *Land Electronic Defence System LEDS-50 MK2*, http://saab.com/land/force-protection/self-protection/leds.
[19] Raytheon, *Active Protection System (APS),* http://www.raytheon.com/capabilities/products/aps.
[20] Rafael Advanced Defense Systems, *Trophy Situational Awareness Situational Awareness and Active Protection Systems* http://www.rafael.co.il/Marketing/349-963-en/Marketing.aspx.
[21] State Company "UKROBORONSERVICE", *The active protection system ZASLON*, http://en.uos.ua/produktsiya/sistemi-zashchiti/49-kompleks-aktivnoy-zashchiti-zaslon.

suggested that the systems are also capable of both selecting and attacking targets autonomously.[22] Illustrative examples include:

- ***aEgis I and II*** *and* ***Super aEgis I and II*** *(Republic of Korea)* are "combat robots" that use various optical, thermal, and infrared sensors to select human targets. They are capable of selecting and attacking targets either by remote operation or autonomously, and are fitted with a range of different weapons. A mobile version, ***Athena***, combines the aEgis weapon with a wheeled vehicle.[23]
- ***Guardium*** *(Israel)* is a mobile ground-combat weapon system that navigates autonomously and can be fitted with various remotely operated weapon systems,[24] although a "fully autonomous" combat version is apparently planned.[25]
- ***MDARS-E (Mobile Detection Assessment and Response System – Exterior)*** *(USA)* is a mobile ground robot that has sensors for detecting humans, and can be fitted with weapons.[26]
- ***Sentry Tech Stationary Remote-Controlled Weapon Station*** *(Israel)* combines various sensors to select human targets with a Samson Remote-Controlled Weapon System.[27] The system can select human targets which are then attacked under remote control by a human operator.[28]

3.4 Sensor-fused munitions, missiles and loitering munitions

Various types of munitions have some level of autonomy in target selection and attack. After launch, these munitions use active sensors, such as radar, to select and attack a target within a designated area. Their on-board automatic target recognition software may also incorporate pre-programmed "signatures" of target objects. Since these types of munitions use on-board guidance systems and do not require further external guidance after launch, they are often described as "fire-and-forget" munitions. Some have been categorized as "semi-autonomous weapons" since they are launched to a specific target area by a human operator before on-board systems for target selection and attack are activated.[29] Other munitions, in particular loitering munitions, have greater autonomy to select and attack targets over a wide area.

Sensor-fused munitions employ on-board active sensor systems for target selection and attack after they have been launched to a specific target area. Illustrative examples include:

- ***BONUS 155 mm projectile*** (*Sweden/France*);[30]
- ***M982 Excalibur 155 mm projectile [Increment III]*** *(USA/Sweden)*;[31]
- ***SMArt 155 mm projectile*** *(Germany);*[32]

[22] S Parkin (2015) Killer robots: The soldiers that never sleep, *BBC*, 16 July 2015, http://www.bbc.com/future/story/20150715-killer-robots-the-soldiers-that-never-sleep; N Shactman (2007) *Robo-Snipers, Auto-Kill Zone to protect Israeli Borders*, *Wired.com,* 4 June 2007, http://www.wired.com/2007/06/for_years_and_y.
[23] DoDamm Systems*, aEgis I & II, Super aEgis I & II & Athena*, http://www.dodaam.com/eng/sub2/menu2.php;
Other systems include the Samsung Techwin SGR-1**:** J Antal, Defending the Republic of Korea, *Military Technology*, October 2010, http://www.stripes.com/machine-gun-toting-robots-deployed-on-dmz-1.110809.
[24] G-nius unmanned ground systems, *Guardium*, http://g-nius.co.il/unmanned-ground-systems/guardium-ugv.html.
[25] G-nius unmanned ground systems, *Guardium* MK III, http://g-nius.co.il/unmanned-ground-systems/guardium-mk-iii.html.
[26] US Navy SPAWAR, *Mobile Detection Assessment and Response System (MDARS) II*, http://www.public.navy.mil/spawar/Pacific/Robotics/Pages/MDARS2.aspx; D Bullock (2008) Gallery: Inside the Navy's Armed-Robot Labs, *Wired*, 2 March 2008, http://www.wired.com/2008/03/gallery-spawar.
[27] Rafael Advanced Defense Systems, *Remote Weapon Stations for Armored Vehicles*, http://www.rafael.co.il/Marketing/402-994-en/Marketing.aspx.
[28] Rafael Advanced Defense Systems, *Sentry Tech*, http://www.rafael.co.il/Marketing/396-1687-en/Marketing.aspx.
[29] For example: US Department of Defense (2012) *Autonomy in Weapon Systems, Directive 3000.09,* (footnote 6), p 14.
[30] BAE Systems, *BONUS 155mm*, http://www.baesystems.com/en/product/155-bonus.
[31] US Army (2007) *Excalibur XM982 Precision Engagement Projectiles*, http://www.dote.osd.mil/pub/reports/FY2007/pdf/army/2007excalibur.pdf.
[32] Gesellschaft für Intelligente Wirksysteme mbH (GIWS), *SMArt 155mm*, http://www.giws.de/en/smart/system.html

- ***STRIX 120 mm*** *mortar round (Sweden).*[33]

Missiles by definition have on-board guidance systems, and some incorporate active sensors to select and attack targets after launch. Generally these types of missiles fly to a pre-programmed or designated location after which they use inbuilt sensors, such as active radar, and information-processing capabilities, such as automatic target-recognition software and pre-programmed signatures of target objects, to determine their target. Some self-destruct or deactivate if a target is not found. Illustrative examples include:

- Air-launched missiles
    - ***AIM-120 Advanced Medium-Range Air-to-Air Missile*** *(USA);*[34]
    - ***Brimstone air-to-surface missile*** *(UK);*[35]
    - ***R-77 / RVV-AE air-to-air missile*** *(Russia).*[36]
- Cruise missiles
    - ***AGM-158 Joint Air-to-Surface Standoff Missile*** *(USA);*[37]
    - ***BrahMos cruise missile*** *(India/Russia).*[38]
- Anti-ship missiles
    - ***C-802 anti-ship missile*** *(China);*[39]
    - ***Long Range Anti-Ship Missile*** *(USA);*[40]
    - ***Naval Strike Missile*** *(Norway).*[41]

Loitering munitions have greater autonomy in selecting and attacking targets when compared to sensor-fused munitions and missiles. Somewhere between a missile and an unmanned air system, these expendable weapon systems select and attack targets over a designated area and period, using on-board sensors and pre-programmed target signatures. Some developmental systems have been designed to carry out attacks as swarms of multiple loitering munitions. Illustrative examples include:

- ***Harpy*** *(Israel)*, an anti-radar weapon that uses on-board sensors and pre-programmed target "signatures" to select and attack radar targets.[42] A related system is the ***Harop***, which can select and attack targets either autonomously or by remote control.[43]
- ***Low-Cost Autonomous Attack System*** *(USA)*, a former developmental system of multiple loitering munitions in swarm that used on-board sensors and pre-programmed target signatures to select and attack targets.[44]
- ***Switchblade*** *(USA)*, a small loitering munition that that uses on-board sensors to select and attack targets, and can operate either autonomously or by remote control.[45]

---

[33] Defense Update, *Strix Precision Guided 120mm Mortar Launched Weapon*, 27 January 2005, http://defense-update.com/20050127_strix.html.
[34] Raytheon, *Advanced Medium-Range Air-to-Air Missile*, http://www.raytheon.com/capabilities/products/amraam
[35] UK Royal Air Force, *Brimstone*, http://www.raf.mod.uk/equipment/brimstone.cfm.
[36] Tactical Missiles Corporation, *RVV-AE Air-to-Air Guided Missile*, http://eng.ktrv.ru/production_eng/323/503/505.
[37] Lockheed Martin, *Joint Air-to-Surface Standoff Missile*, http://www.lockheedmartin.co.uk/us/products/jassm.html.
[38] K Mizokami (2015) Bullseye: The 5 Most Deadly Anti-Ship Missiles of All Time, *National Interest*, 13 March 2015, http://nationalinterest.org/feature/bull%E2%80%99s-eye-the-5-most-deadly-anti-ship-missiles-all-time-1241.
[39] C Carlson (2013) China's Eagle Strike-Eight Anti-Ship Cruise Missiles: YJ-81, YJ-82, and C802, *Defense Media Network*, 6 February 2013, http://www.defensemedianetwork.com/stories/chinas-eagle-strike-eight-anti-ship-cruise-missiles-yj-81-yj-82-and-c802/.
[40] Lockheed Martin, *Long Range Anti-Ship Missile*, http://www.lockheedmartin.com/us/products/LRASM.html.
[41] Kongsberg Defence Systems, *Naval Strike Missile – NSM*, http://www.kongsberg.com/en/kds/products/missilesystems/navalstrikemissile.
[42] Rafael Advanced Defense Systems, *HARPY NG*, http://www.iai.co.il/2013/36694-16153-en/IAI.aspx.
[43] Rafael Advanced Defense Systems, *HAROP*, http://www.iai.co.il/2013/36694-46079-en/Business_Areas_Land.aspx.
[44] Defense Update, *Low Cost Autonomous Attack System - Lockheed Martin*, 26 July 2006, http://defense-update.com/products/l/locaas.htm.

- ***Tactical Advanced Recce Strike (TARES) unmanned combat air vehicle*** *(Germany)*, a developmental loitering munition that uses on-board sensors and pre-programmed target signatures to select and attack targets.[46]

3.5 Torpedoes and encapsulated torpedo mines

Similar to some missiles (Section 3.4), certain torpedoes incorporate on-board active acoustic sensors to select and attack targets after launch. Other underwater weapon systems with significant autonomy include encapsulated torpedo mines. These sea mines also use active acoustic sensors to detect ships or submarines. Once they are detected, the weapon system selects the target and launches an attack with a torpedo.[47] Illustrative examples include:

- ***MK 48 Heavyweight Torpedo*** *(Australia/USA)* is an anti-submarine and anti-surface warfare torpedo that can either be wire-guided or can use on-board active sensors to select and attack targets after launch.[48]
- ***MK 60 CAPTOR Encapsulated Torpedo*** *(USA)* is a sea mine (no longer in service), tethered to the sea floor, that is activated when it detects a target submarine and then fires a torpedo to attack it.[49]
- ***MU90/IMPACT Advanced Lightweight Torpedo*** *(France/Italy)* is an anti-submarine torpedo that is pre-programmed before launch and then uses on-board active sensors to select and attack the target.[50]
- ***PMK-2 encapsulated torpedo mine*** *(China/Russia)* is a sea mine that is activated when it detects a target ship and then fires a torpedo to attack it.[51]

[45] Aerovironment, *Switchblade*, https://www.avinc.com/uas/adc/switchblade.
[46] Army Technology, *TARES Unmanned Combat Air Vehicle (UCAV) - Rheinmetall Defence Electronics*, http://www.army-technology.com/projects/taifun.
[47] P Scharre and M Horowitz (2015) *An Introduction to Autonomy in Weapon Systems*, (footnote 6).
[48] Lockheed Martin, *Mk 48 MOD 7 Common Broadband Advanced Sonar Systems (CBASS) Heavyweight Torpedo*, http://www.lockheedmartin.com/content/dam/lockheed/data/ms2/documents/CBASS-brochure.pdf; US Navy (2013) Fact File: MK 48 - Heavyweight Torpedo, 16 December 2013, http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=950&ct=2.
[49] Federation of American Scientists, *MK 60 Encapsulated Torpedo (CAPTOR)*, 13 December 1998, http://fas.org/man/dod-101/sys/dumb/mk60.htm.
[50] Eurotorp, *MU 90/IMPACT*, http://www.eurotorp.com/the-products/mu90-impact_25.
[51] *Ibid.*

4. EMERGING TECHNOLOGY AND FUTURE AUTONOMOUS WEAPON SYSTEMS

Existing weapons with autonomy in their critical functions of selecting and attacking targets have tended to be quite constrained in their functions and the contexts in which they are used. Future autonomous weapon systems, however, might be given more freedom of action, for example: increased mobility to operate outside tightly constrained spatial and temporal limits, and increased capacity to determine their own functions and targets, or to react to changing circumstances.

Among many technologically advanced militaries, there has been a dramatic increase during the past 15 years in the use of robotic unmanned systems in the air, on land and at sea, and increasingly they are being adapted, designed and used as weapon systems.[52] Proliferation has been rapid and widespread,[53] and – although most existing robotic weapon systems (such as armed drones) are operated by remote control for the critical functions of selecting and attacking targets – there is an overall trend towards increasing autonomy of these systems.[54]

The important point is that the architecture for possible future autonomous weapon systems is starting to take shape in at least two forms: firstly, remote-controlled robotic weapon systems that could be adapted to select and attack targets autonomously; and secondly, commercial (non-weaponized) autonomous robotic systems that could be adapted as weapon systems.

These two routes to future autonomous weapon systems are not difficult to envisage. The first might only require a software upgrade, especially for weapon systems that can already select targets autonomously while still requiring human authorization to attack (e.g. the anti-personnel "sentry" weapons discussed in Section 3.3). A change in software could close this targeting loop, allowing the machine to select and attack targets autonomously.

The second route is perhaps more difficult to predict, but commercial and academic scientific advances have already produced robotic systems that can navigate and fly autonomously, and even operate as part of an interconnected swarm.[55] Many of these systems are small, cheap and accessible, and could, in theory, be weaponized by States or non-State armed groups.

The range of possible future autonomous weapon systems, therefore, is broad, and could include weaponized robotic systems already envisaged for operations in the air, on the ground, and at sea. In the air, these systems might be used for air-to-air combat and suppression of air defences, in addition to their current use for targeted attacks. Some of these weapon systems already have a level of autonomy for aspects of target selection, but

---

[52] P W Singer (2009) *Wired for War*, Penguin, New York; US Department of Defense (2013) *Unmanned Systems Integrated Roadmap FY2013-2038*; P Rogers (2014) Unmanned Air Systems: The Future of Air & Sea Power? *Institut Français des Relations Internationales (IFRI) Focus Stratégique,* No. 49, January 2014 ; P Scharre (2014) *Robotics on the Battlefield Part I: Range, Persistence and Daring*, Center for a New American Security, May 2014; P Scharre (2014) *Robotics on the Battlefield Part II: The Coming Swarm*, Center for a New American Security, October 2014.

[53] S Joshi and A Stein (2013) Emerging Drone Nations. *Survival*, Vol. 55:5, pp. 53–78; R O'Gorman and C Abbott (2013) *Remote control war: Unmanned combat air vehicles in China, India, Israel, Iran, Russia and Turkey*, Open Briefing, 20 September 2013; M Zenko and S Kreps (2014) *Limiting Armed Drone Proliferation*, Council on Foreign Relations, Special Report No 69, June 2014; K Sayler (2015) *A World of Proliferated Drones: A Technology Primer*, Center for a New American Security, June 2015.

[54] US Department of Defense, Defense Science Board (2012) *Task Force Report: The Role of Autonomy in DoD Systems,* 19 July 2012; NATO (2014) *Policy Guidance: Autonomy in Defence Systems, Multinational Capability Development Campaign (MCDC) 2013-2014, Focus Area "Role of Autonomous Systems in Gaining Operational Access"*, NATO Supreme Allied Commander Transformation HQ, 29 October 2014; D Gonzales and S Harting (2014) *Designing Unmanned Systems with Greater Autonomy: Using a Federated, Partially Open Systems Architecture Approach*, RAND Corporation; P Scharre (2014) *Robotics on the Battlefield Part II: The Coming Swarm*, (footnote 52).

[55] P Scharre (2014) *Robotics on the Battlefield Part II: The Coming Swarm*, (footnote 52); D Hambling (2016) Drone swarms will change the face of modern warfare, *Wired*, 7 January 2016, http://www.wired.co.uk/news/archive/2016-01/08/drone-swarms-change-warfare.

future systems could be increasingly autonomous.[56] A large range of weaponized systems are being developed, from small quadcopters[57] to those the size of manned aircraft.[58]

On-the-ground robotic weapon systems, whether stationary or mobile, might be used for defence of perimeters and borders, in military combat operations,[59] and in armed law-enforcement operations.[60] A range of robotic ground systems[61] has already been tested with weapons to enable, at minimum, remote operation.[62] However, there is already discussion of moving from remote-control targeting to "supervised autonomy".[63]

At sea, robotic maritime systems of various sizes and functions are being developed as weapon platforms, including unmanned surface vehicles,[64] which might be used for anti-submarine and surface warfare, and could operate as swarms,[65] as well as unmanned underwater vehicles,[66] which might be used for anti-submarine warfare, laying mines and other types of attack.[67] Autonomous capability for underwater weapon systems is of particular military interest owing to the difficulties of communication underwater and the size of potential operating areas.[68]

With these military technical developments in mind, the following section will summarize some of the legal and ethical questions raised by increasing autonomy in the critical functions of weapon systems.

---

[56] P W Singer (2013) *The Predator Comes Home: A Primer on Domestic Drones, their Huge Business Opportunities, and their Deep Political, Moral, and Legal Challenges*, *Brookings Institution*, 8 March 2013; P Scharre (2014) *Robotics on the Battlefield Part I: Range, Persistence and Daring*, (footnote 52).

[57] For example: Israel Aerospace Industries, *ROTEM Tactical Loitering Munition*, http://www.iai.co.il/2013/36694-46735-en/Business_Areas_Land.aspx; Russia Today (2106) Killer drone squad: Russia unveils anti-armor assault multicopter, 10 February 2016, *Russia Today*, https://www.rt.com/news/332045-tank-destroyer-drone-complex.

[58] For example: Northrop Grumman, *X-47B*, http://www.northropgrumman.com/Capabilities/x47bucas/Pages/default.aspx; BAE Systems, *Taranis*, http://www.baesystems.com/en/product/taranis; Dassault Aviation, *nEUROn*, http://www.dassault-aviation.com/en/defense/neuron/introduction.

[59] P McLeary (2014) US army Studying Replacing Thousands of Grunts with Robots, *Defense News,* 20 January 2014, http://www.defensenews.com/article/20140120/DEFREG02/301200035/US-Army-Studying-Replacing-Thousands-Grunts-Robots.

[60] For example ground systems employing electrical weapons or 'tear gas': M Jewell (2007) Taser, IRobot Team Up to Arm Robots, *Washington Post*, 1 July 2007, http://www.washingtonpost.com/wp-dyn/content/article/2007/07/01/AR2007070101096.html; M Crowley (2015) *Tear Gassing by Remote Control*, Oxford Research Group, December 2015.

[61] For example: QinetiQ, *Talon SWORDS*, http://www.qinetiq.com/media/news/releases/Pages/talon-robots-demo-swords-at-dsei.aspx; QinetiQ, *Modular Advanced Armed Robotic System*, https://www.qinetiq-na.com/products/unmanned-systems/maars; Northrop Grumman, *Andros*, http://www.northropgrumman.com/Capabilities/Remotec/Applications/Pages/Swat.aspx.

[62] J Bloom (2008) Robots ready to support soldiers on the battlefield, *Guardian*, 26 June 2008, http://www.theguardian.com/technology/2008/jun/26/robots.weaponstechnology; Army Times (2013) UGV Models Face Off Over Firepower, Load Carrying, *Army Times*, 12 October 2013; Moscow Times (2015) Russian Battle Robots Near Testing for Military Use, *Moscow Times*, 2 January 2015, http://www.themoscowtimes.com/news/article/russian-battle-robots-near-testing-for-military-use/514038.html..

[63] E Lopes (2014) *Remote lethality: Army researchers address a host of challenges*, US Army, Picatinny Arsenal Public Affairs, 10 November 2014, http://www.army.mil/article/137493/Remote_lethality__Army_researchers_address_a_host_of_challenges/

[64] For example: Rafael Defense Systems*, Protector Unmanned Naval Patrol Vehicle*, http://www.rafael.co.il/Marketing/288-1037-en/Marketing.aspx.

[65] P Tucker (2014) Inside the Navy's Secret Swarm Robot Experiment, *Defense One*, 5 October 2014, http://www.defenseone.com/technology/2014/10/inside-navys-secret-swarm-robot-experiment/95813.

[66] For example: DARPA, *Anti-Submarine Warfare Continuous Trail Unmanned Vessel*, http://www.darpa.mil/program/anti-submarine-warfare-continuous-trail-unmanned-vessel.

[67] A Martin (2012) U.S. Expands Use Of Underwater Unmanned Vehicles, *National Defense*, April 2012, http://www.nationaldefensemagazine.org/archive/2012/April/Pages/USExpandsUseOfUnderwaterUnmannedVehicles.aspx; N. Hopkins (2012) Ministry of Defence plans new wave of unmanned marine drones, *Guardian*, 2 August 2012, http://www.theguardian.com/world/2012/aug/02/ministry-defence-plans-unmanned-marine-drones.

[68] United Nations Institute for Disarmament Research (UNIDIR) (2015), *The Weaponization of Increasingly Autonomous Technologies in the Maritime Environment: Testing the Waters*, UNIDIR Resources No 4, 2015.

5. LEGAL AND ETHICAL IMPLICATIONS OF INCREASING AUTONOMY[69]

Although new technologies of warfare, including autonomous weapon systems, are not specifically regulated by international humanitarian law (IHL) treaties, their development and employment in armed conflict does not occur in a legal vacuum. As with all weapon systems, they must be capable of being used in compliance with IHL, and in particular with its rules on the conduct of hostilities. The responsibility for ensuring this rests, first and foremost, with each State that is developing and deploying these new weapons.

While it is undisputed that autonomous weapon systems must be capable of being used in accordance with IHL, difficulties in interpreting and applying these rules to new weapons may arise in view of their unique characteristics, the intended and expected circumstances of their use, and their foreseeable consequences in humanitarian terms.

Ultimately, these difficulties may raise the question of whether the existing law is sufficiently clear, or whether there is a need to clarify IHL and improve interpretive standards or to develop new rules.

5.1 Compliance with international humanitarian law

There is no doubt that the IHL rules on the conduct of hostilities are addressed to human beings. While the primary subjects of IHL are States, rules on the conduct of hostilities (notably the rules of distinction, proportionality and precautions in attack) are addressed to those who plan and decide upon an attack. These rules create obligations for human combatants and fighters, who are responsible for respecting them and will be held accountable for violations.

A fundamental legal question is, therefore, whether compliance with some or all IHL rules on the conduct of hostilities inherently requires human judgement. If it is concluded that there is such an inherent requirement under IHL, then the question arises as to whether this requirement can be satisfied by human involvement in the development and deployment stages of the "life cycle" of an autonomous weapon system (see Section 6.2). If so, the overall question remains of whether a particular autonomous weapon system could, in the intended and expected circumstances of its use, function in a way that complies with the relevant rules of IHL.

Based on current and foreseeable technology, ensuring that autonomous weapon systems can be used in compliance with IHL will pose a formidable challenge if these weapons were to be used for more complex tasks or deployed in more dynamic environments than has been the case until now.

Key questions, closely linked to the context in which an autonomous weapon system is used, include whether the weapon 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 weapon system would function in a way that respects the obligation to weigh up 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 weapon 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

[69] Parts of this section draw from: ICRC (2015) *International humanitarian law and the challenges of contemporary armed conflicts* (footnote 5).

be expected to violate the rule of proportionality, as required by the rules on precautions in attack.

For autonomous weapon 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. These are inherently qualitative assessments in which unique human reasoning and judgement will continue to be required.

In view of these realities, there remain serious doubts about the capability of developing and using autonomous weapon systems that would comply with IHL in all but the narrowest of scenarios and the simplest of environments, at least for the foreseeable future. In this respect, it seems evident that overall human control over the selection and attack of targets will continue to be required to ensure respect for IHL.

The ability to exert human control over selecting and attacking targets, and thereby ensure compliance with IHL, will depend on at least two aspects. Firstly, the **technical characteristics and performance of the autonomous weapon system** (through its sensors, computer programming, and weapon(s)) will be important, in particular its predictability and reliability, which in turn may be influenced by the degree to which the weapon system can determine its own functions or set its own goals.

Secondly, the **operational parameters** for use of the weapon system, including the circumstances in which the weapon is intended to be used and the constraints (or lack thereof) placed on its freedom of operation, will also have a significant impact on the ability to exert human control. Important operational parameters may include:[70]

- the task the weapon system is being used for (e.g. offensive or defensive operations);
- the type of target it attacks (e.g. objects and/or personnel);
- the type of force it is using (e.g. kinetic or non-kinetic) and type of munitions (e.g. bullets or explosive weapons);
- the environment in which it is used (e.g. air, ground, or sea; simple or “cluttered” environments);
- the mobility of the weapon in space (e.g. stationary or mobile; narrow or wide geographical area);
- the time frame of the action of the weapon (e.g. short or long periods); and
- the level of human supervision of the weapon (e.g. supervised or unsupervised; ability to deactivate or not).

The combination of the technical characteristics and performance of the weapon system with the operational parameters of its use are critical to determining the foreseeable effects of a particular weapon, and therefore in determining whether it can be used in conformity with IHL rules.

5.2 Legal review

In accordance with Article 36 of Additional Protocol I, 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

[70] ICRC (2015) *Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects*, (footnote 2), p 15.

international law. Legal reviews of new weapons, including autonomous weapon systems, are a critical measure for States to ensure respect for IHL.[71]

The above challenges for IHL compliance will need to be carefully considered by States when carrying out legal reviews of any autonomous weapon 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 weapon systems were to become more complex or to be given more freedom of action in their operations, and therefore become less predictable.

Questions arise as to how "targeting rules" (e.g. the rules of proportionality and precautions in attack) are considered in reviewing weapons. 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.

Generally, the lawfulness of a particular attack is the responsibility of the operational commander who selects and decides to attack a target, applying IHL targeting rules with the assistance of his/her legal adviser. Thus, the deployment of an autonomous weapon system, especially one given more freedom of action in time, space and operations, might raise questions about whether the commander, in deploying the weapon, can ensure compliance with those rules.

Predictability about the operation of an autonomous weapon system in the context in which it is to be used must be sufficiently high to allow an accurate legal review. Indeed, deploying a weapon 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 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 weapon 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 are 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 in assessing whether an autonomous weapon system will perform as anticipated in the intended or expected circumstances of use.

5.3 Accountability

Some have raised concerns that the use of autonomous weapon systems may lead to an "accountability gap" in case of violations of IHL. Others are of the view that no such gap would ever exist as there will always be a human involved in the decision to deploy the weapon to whom responsibility could be attributed. Still, it is unclear how responsibility could be attributed in relation to unpredictable operations of autonomous weapon systems.

For instance, under IHL and international criminal law, the limits of control over, or the unpredictability of, an autonomous weapon system could make it difficult to find individuals involved in the programming and deployment of the weapon liable for serious violations of

[71] ICRC (2006) *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*, Geneva, January 2006, www.icrc.org/eng/assets/files/other/icrc_002_0902.pdf.

IHL. They may not have the knowledge or intent required for such a finding, owing to the fact that the machine can select and attack targets independently.

Programmers might not have knowledge of the concrete situations in which, at a later stage, the weapon might be deployed and IHL violations could occur. On the other hand, a programmer who intentionally programs an autonomous weapon to commit war crimes would certainly be criminally liable. Likewise, a commander would be liable for deciding to use an autonomous weapon system in an unlawful manner, for example, deploying in a populated area an anti-personnel autonomous weapon that is incapable of distinguishing civilians from combatants. In addition, a commander who knowingly decides to deploy an autonomous weapon whose performance and effects he/she cannot predict may be held criminally responsible for any serious violations of IHL that ensue, to the extent that his/her decision to deploy the weapon is deemed reckless under the circumstances.

Under the law of State responsibility, in addition to accountability for violations of IHL committed by members of its armed forces, a State could be held liable for violations of IHL resulting from the use of an autonomous weapon system that it has not, or has inadequately, tested or reviewed prior to deployment. Further, under the laws of product liability, manufacturers and programmers could also be held accountable for errors in programming or for the malfunction of an autonomous weapon system.

5.4 Ethical considerations

Autonomous weapon systems also raise ethical concerns that deserve careful consideration. The fundamental question at the heart of these concerns, and irrespective of whether the weapon systems can be used in compliance with IHL, is whether the principles of humanity and the dictates of public conscience would allow machines to make life-and-death decisions in armed conflict without human involvement.[72] The debates of recent years among States, experts, civil society and the public have shown that there is a sense of deep discomfort with the idea of any weapon system that places the use of force beyond human control. The question remains, however: what degree of human control is required, and in which circumstances, in light of ethical considerations? Is it sufficient for a human being to program an autonomous weapon system according to certain parameters and then make the decision to deploy it in a particular context? Or is it necessary for a human being also to bring his or her judgement to bear on each individual attack? If the weapon autonomously uses force against a human target, what ethical considerations would this entail?

With increasing autonomy in the critical functions of weapon systems, a point may be reached where humans are so far removed in time and space from the selection and attack of targets that human decision-making regarding the use of force is substituted with machine processes. This raises profound moral and societal questions about the role and responsibility of humans in the use of force and the taking of human life. It is this possibility that would represent a paradigm shift in the conduct of hostilities, which until now has been exclusively or primarily controlled by human beings.

Since weapon systems that operate with varying degrees of autonomy in their critical functions already exist, and new advances are constantly emerging, there is a need to approach these issues with a sense of responsibility and urgency so that we do not allow technological advances to outpace our ethical deliberations.

[72] The "principles of humanity and the dictates of public conscience" are mentioned notably in Article 1(2) of Additional Protocol I and in the preamble of Additional Protocol II to the Geneva Conventions, referred to as the Martens Clause. In its 1996 Advisory Opinion on the legality of the threat or use of nuclear weapons, the International Court of Justice (ICJ) has affirmed that the applicability of the Martens Clause "is not to be doubted" (para. 87) and that it has "proved to be an effective means of addressing the rapid evolution of military technology" (para. 78); ICJ (1996) *Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion,* ICJ Reports 1996.

6. HUMAN CONTROL

6.1 Possible elements of human control

There is broad agreement that "meaningful", "appropriate" or "effective" **human control** over weapon systems and the use of force must be retained, for legal, ethical and/or policy reasons. Different terms are used by a range of stakeholders, including experts and State representatives.[73] Although the term "meaningful human control"[74] has been raised more frequently as a starting point for further discussions, all those terms can be considered, at this early stage of the debate, as synonymous.[75] The discussion of human control stems from the general assumption that no weapon system can be permitted to operate entirely beyond human control.

There are three main drivers underpinning the need for meaningful human control. Firstly, there is the requirement to comply with IHL rules in the conduct of hostilities (see Section 5.1). In this connection, some argue that the absence of meaningful human control in selecting and attacking targets could lead to an accountability gap in cases where the use of autonomous weapon systems resulted in a violation of IHL. The second driver for meaningful human control is embodied in the principles of humanity and the dictates of public conscience, in particular the ethical and moral considerations inherent in decisions to take life.[76] Thirdly, from a military perspective, there is interest in maintaining command and control over weapon systems for operational reasons.

While it is broadly accepted that a weapon system selecting and attacking targets without any human control would be considered unacceptable, questions arise as to the type and degree of human control that would be considered "meaningful" and at which stages of the "life cycle" of the weapon system.

Based on the three core drivers for meaningful human control – IHL compliance, ethical acceptability and military operational requirements – there are a number of important elements that may determine whether human control is meaningful, including:[77]

- **predictability** of the weapon system in its intended or expected circumstances of use;
- **reliability** of the weapon system in its intended or expected circumstances of use;
- **human intervention** in the functioning of the weapon system during its development, deployment and use;
- knowledge and accurate **information** about the functioning of the weapon system and the context of its intended or expected use; and
- **accountability** for the functioning of the weapon system following its use.

---

[73] *Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*, Submitted by the Chairperson of the Informal Meeting of Experts, UN Document, CCW/MSP/2015/3, 2 June 2015.

[74] The term "meaningful human control" was first coined by the NGO Article 36. See: Article 36 (2014) *Key Areas for Debate on Autonomous Weapon Systems*, Briefing Paper, May 2014.

[75] In this section, the term "meaningful human control" will be used.

[76] United Nations Institute for Disarmament Research (UNIDIR) (2014) *The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control Might Move the Discussion Forward*, UNIDIR Resources No. 2, 2014, p. 3.

[77] Article 36 has highlighted reliable information, human action, and accountability as key elements. See: Article 36 (2013) *Killer Robots: UK Government Policy on Fully Autonomous Weapons*, April 2013. See also: M Brehm (2015) *Meaningful Human Control, Presentation to the informal meeting of experts on lethal autonomous weapon systems of the Convention on Certain Conventional Weapons (CCW), Geneva, 14 April 2015*. Horowitz and Scharre refer to three essential components that characterize meaningful human control: "1. Human operators are making informed, conscious decisions about the use of weapons. 2. 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. 3. The weapon is designed and tested, and human operators are properly trained, to ensure effective control over the use of the weapon." See: M Horowitz and P Scharre (2015) *Meaningful Human Control in Weapon Systems: A Primer*, Center for a New American Security, March 2015, pp.14-15. UNIDIR refers to a number of parameters that may shape human control, such as the function of the weapon, the spatial limitations, the time limitations, and predictability. See: UNIDIR (2014) *The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control Might Move the Discussion Forward* (footnote 76), pp. 5-9.

6.2 Human control at different stages

Control exercised by human beings can take various forms and operate at different stages of the "life cycle" of an autonomous weapon system, including: (1) the development of the weapon system, including its programming; (2) the deployment and use of the weapon system, including the decision by the commander or operator to use or activate the weapon system; and (3) the operation of the weapon system, during which it selects and attacks targets.

At the first stage, the humans designing and programming the autonomous weapon system exert some control over the system by ensuring that it is reliable and predictable in its functioning in real-world scenarios. It is at this first stage that there must be a review of the weapon to ensure it can be used in accordance with IHL and other applicable law, and otherwise to determine whether any constraints should apply to the use of the weapon for ethical or policy reasons. Multidisciplinary expertise is important to understanding how the weapon functions, its capabilities, and its limitations.[78] Constraints on or parameters for the use of the weapon can be integrated into the military instructions for the use of the weapon, for instance to limit the use to a specific environment or situation. The review procedure is important to assess the predictability and reliability of the weapon system given its intended or expected circumstances of use.[79]

The second stage at which human control is exerted is during deployment and use of the autonomous weapon system, which involves the commander's decision to use a particular weapon system for a particular purpose in a specific situation. Human control on the part of the commander is based on sufficient knowledge and understanding of the weapon's functioning and proper training to ensure that, when deployed in a specific situation, it operates in accordance with IHL and any other restrictions that apply to its use. The commander must gather adequate information on the area of operation of the weapon, including whether or not civilians or civilian objects are present. The question arises as to what level of information relating to the area of operation would be required to make an informed decision, as well as to the level of training necessary for the military personnel involved in this stage.[80]

It is clear that human control is exerted in the development and deployment stages of the weapon's "life cycle". It is in stage three, however, when the weapon is in operation – when it autonomously selects (i.e. searches for or detects, identifies, tracks, selects) and attacks (i.e. uses force against, neutralizes, damages or destroys) the target(s) – that the important question arises as to whether human control in the first two stages is sufficient to overcome minimal or no human control at this last stage.

In this third stage, the ability to exert human control will depend on the **technical characteristics and performance** of the weapon system in question – in particular the predictability and reliability of its functioning – and the **operational parameters** of its use, including: the task that the weapon system is being used for; the type of target it attacks; the type of force it is using; the environment in which it is used; the mobility of the weapon in space; the time frame of the action of the weapon; and the level of human supervision of the weapon (see Section 5.1).

---

[78] ICRC (2006) *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* (footnote 71).
[79] For example, the U.S. Department of Defense has listed requirements for the design of autonomous weapon to "allow commanders and operators to exercise appropriate levels of human judgment over the use of force" including "hardware and software verification and validation (V&V) and realistic system developmental and operational test and evaluation (T&E). See: US Department of Defense (2012) *Autonomy in Weapon Systems, Directive 3000.09* (footnote 6), p 2.
[80] M Horowitz and P Scharre (2015) *Meaningful Human Control in Weapon Systems: A Primer* (footnote 77), p 14.

Although there is agreement on the importance of maintaining human control over weapon systems and the use force, further discussion is needed of the ways in which this control is exerted, what makes it meaningful, and the requirements from a legal, ethical and military operational perspective. During these discussions, it will be useful to identify how human control will be ensured over the operation of the weapon system, through which processes and constraints, and at which stages of the "life cycle" of the weapon system. Ultimately, these considerations will provide insights into where the limits of autonomy in weapon systems must be placed.

ANNEX 1: EXPERT MEETING PROGRAMME

Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons

Lake Geneva Hotel, Versoix, Switzerland, 15–16 March 2016

DAY ONE – 15 MARCH 2016

9:00 – 9:15 WELCOME AND INTRODUCTION

***Dr Knut DÖRMANN**, Head of the Legal Division and Chief Legal Officer, ICRC*

9:15 – 10:30 SESSION 1: CHARACTERISTICS OF AUTONOMOUS WEAPON SYSTEMS

Chair: *Dr Knut Dörmann, Head of the Legal Division and Chief Legal Officer, ICRC*

Session objective: Discuss the characteristics of autonomous weapon systems and explain the working definition for the meeting.

Key questions:

- *What are the defining characteristics of autonomous weapon systems?*
- *What distinguishes them from other weapons?*
- *What are the "critical functions" of weapon systems that are most relevant?*
- *Is there a difference between automation and autonomy in the critical functions?*
- *What is the working definition of an "autonomous weapon system" for this meeting?*

Speakers: ***Dr Martin HAGSTRÖM**, Swedish Defence Research Agency, Sweden*

***Lt Col. Alan SCHULLER**, Stockton Center for the Study of International Law, US Naval War College, United States*

***Dr Neil DAVISON**, Scientific and Policy Adviser, Arms Unit, Legal Division, ICRC*

Discussion

SESSION 2: AUTONOMY IN EXISTING WEAPONS

Note: The session objective and key questions apply to all four parts of Session 2.

Chair: *Dr Neil Davison, Scientific and Policy Adviser, Arms Unit, Legal Division, ICRC*

Session objective: Understand autonomy in the critical functions of existing weapon systems.

Key questions:

- *Which existing weapons have autonomy in the critical functions of selecting and attacking targets?*
- *How do these weapons function? In which roles? Against which targets? In which situations and environments? For what time periods? Over which areas?*
- *What are their technical characteristics and performance? (e.g. predictability, reliability, identification of targets)?*
- *What are the operational parameters placed on their use, and the reasons for these constraints (e.g. tasks, targets, operating environments, mobility, time frame of action, human supervision)?*
- *How is the required level of human control over the functioning of these weapons determined and applied in practice?*

11:00 – 12:30	**SESSION 2.1: MISSILE- AND ROCKET-DEFENCE WEAPONS**
Speakers:	***Dr Brian HALL**, Joint Chiefs of Staff, Department of Defense, United States*
	***Gp Capt. Ajey LELE (Ret'd)**, Institute for Defence Studies and Analyses, India*
	Discussion
14:00 – 15:00	**SESSION 2.2: VEHICLE "ACTIVE-PROTECTION" WEAPONS AND ANTI-PERSONNEL "SENTRY" WEAPONS**
Speaker:	***Dr Gabi SIBONI**, Institute for National Security Studies, Israel*
	Discussion
15:00 – 16:00	**SESSION 2.3: SENSOR-FUSED MUNITIONS, MISSILES AND LOITERING MUNITIONS**
Speaker:	***Dr Heather ROFF**, Global Security Initiative, Arizona State University, United States*
	Discussion
16:30 – 17:30	**SESSION 2.4: TORPEDOES AND ENCAPSULATED TORPEDO MINES**
Speaker:	***Commander Matthias ELVERT**, German Naval Command, Ministry of Defence, Germany*
	Discussion

DAY TWO – 16 MARCH 2016

09:00 – 10:30	SESSION 3: EMERGING TECHNOLOGY AND FUTURE AUTONOMOUS WEAPONS

Chair: *Dr Gilles Giacca, Legal Adviser, Arms Unit, Legal Division, ICRC*

Session objective: Explore emerging technology and the implications for future autonomous weapon systems.

Key questions:

- *What is the range of weaponized military robotic systems (air, land and sea) under development?*
- *Which of these systems might in the future have autonomy in their critical functions of selecting and attacking targets?*
- *What is the range of civilian autonomous robotic systems that might in the future be weaponized?*
- *How might future autonomous weapon systems differ from existing weapons (e.g. mobility, functions, goal-setting, adaptability or machine learning, size, numbers or swarms)?*

Speakers: ***Dr Markus HOEPFLINGER**, Armasuisse, Federal Department of Defence, Civil Protection and Sport, Switzerland*

***Dr Albert EFIMOV**, Skolkovo Robotics Center, Russian Federation*

Discussant: ***Dr Ludovic RIGHETTI**, Max Planck Institute for Intelligent Systems, Germany*

Discussion

11:00 – 12:30	SESSION 4: LEGAL AND ETHICAL IMPLICATIONS OF INCREASING AUTONOMY

Chair: *Ms Kathleen Lawand, Head of the Arms Unit, Legal Division, ICRC*

Session objective: Examine the legal and ethical implications of increasing autonomy in weapon systems.

Key questions:

- *Which factors are most relevant for ensuring compliance with international humanitarian law (IHL)?*
- *How important are the technical characteristics and performance of the weapon (e.g. predictability and reliability)?*
- *How important are the operational parameters governing use of the weapon (e.g. tasks, targets, operating environments, mobility, time frame of action, human supervision)?*
- *Which future developments might create challenges for IHL compliance or for accountability?*
- *Which future developments might raise questions of ethical acceptability?*
- *What mechanisms are in place to monitor research and development of new weapons? Is there specific guidance for legal reviews of autonomous weapon systems?*

Speakers: ***Col. ZHANG Xinli**, Ministry of Defence, China*

***Dr Gilles GIACCA**, Legal Adviser, Arms Unit, Legal Division, ICRC*

***Prof**. **Paola GAETA**, The Graduate Institute, Switzerland*

Discussion

14:00 – 15:30	**SESSION 5: HUMAN CONTROL**

Chair: *Dr Gilles Giacca, Legal Adviser, Arms Unit, Legal Division, ICRC*

Session objective: Consider the legal, military (operational) and ethical requirements for human control over weapon systems and the use of force.

Key questions:

- *What is understood by meaningful, appropriate, or effective human control over weapon systems and the use of force?*
- *To what extent does IHL require human control over targeting decisions in armed conflict?*
- *What are the military operational requirements for human control?*
- *How is the notion of human control informed by ethical considerations?*
- *What are the key elements needed to ensure human control over weapon systems and the use of force (e.g. predictability, reliability, human intervention, information, accountability mechanisms)?*
- *How is human control exerted at different stages of the research and development, deployment and use of autonomous weapon systems?*
- *Does the nature of human control required vary according to the context (e.g. specific weapon and intended or expected circumstances of use)?*

Speakers: ***Mr Richard MOYES**, Article 36, United Kingdom*

***Ms Merel EKELHOF**, VU University Amsterdam, Netherlands*

***Dr Peter ASARO,** School of Media Studies, The New School, United States*

Discussion

16:00 – 18:00	**SESSION 6: ADDRESSING THE CHALLENGES RAISED BY INCREASING AUTONOMY**

Chair: *Ms Kathleen Lawand, Head of the Arms Unit, Legal Division, ICRC*

Session objective: Discuss possible approaches to addressing the challenges raised by increasing autonomy.

Key questions:

- *With increasing autonomy, how will States ensure human control over weapon systems and the use of force?*
- *Is there a need to develop specific requirements in terms of human control? If so, how might these be developed?*
- *Is there a need to develop specific limits on autonomy in weapons to ensure compliance with IHL and ethical acceptability? If so, where might these limits lie?*
- *What are the lessons from existing weapons for considering human control or limits on autonomy?*
- *What is the role for national legal review processes under Article 36 of Additional Protocol I?*

Speakers: ***Lt Col. John STROUD-TURP**, Ministry of Defence, United Kingdom*

***Dr Vadim KOZYULIN**, PIR Center for Policy Studies, Russian Federation*

***Ms Kerstin VIGNARD,** United Nations Institute for Disarmament Research*

Discussion

Closing remarks by the Chair and end of the meeting

+ + +

ANNEX 2: LIST OF PARTICIPANTS

GOVERNMENT EXPERTS (* indicates speaker)

Algeria	Mr Nader LOUAFI First Secretary, Permanent Mission of Algeria to the United Nations and other international organizations in Geneva
Australia	Mr Hugh WATSON First Secretary, Legal Adviser, Permanent Mission of Australia to the Conference on Disarmament, Geneva
Brazil	Mr Diogo RAMOS COELHO Third Secretary, Division of Disarmament and Sensitive Technologies, Ministry of Foreign Affairs of Brazil
	Mr Carlos Guilherme SAMPAIO FERNANDES Second Secretary, Permanent Mission of Brazil to the Conference on Disarmament, Geneva
China	Mr DAI Bin Expert, Ministry of Defence
	Mr HU Yi Official, Ministry of Defence
	Ms YANG Jia Second Secretary, Ministry of Foreign Affairs
	Col. ZHANG Xinli * Expert, Ministry of Defence
Egypt	Mr Mohamed GATTA AMAL Second Secretary, Permanent Mission of Egypt to the United Nations and other international organizations in Geneva
France	Mr Dany BURIGANA Colonel (Armaments), Counter-Proliferation and Arms Transfers Control, Directorate-General for International Relations and Strategy, Ministry of Defence
	H.E. Ms Alice GUITTON Ambassador, Permanent Representative, Permanent Mission of France to the Conference on Disarmament, Geneva
	Col. Nicolas COUSSIERE Military Adviser, Permanent Mission of France to the Conference on Disarmament, Geneva
	Ms Marie-Gaelle ROBLES Counsellor, Permanent Mission of France to the Conference on Disarmament, Geneva

Germany	H.E. Mr Michael BIONTINO Ambassador, Permanent Mission of Germany to the Conference on Disarmament, Geneva
	Mr Matthias ELVERT * Commander, German Naval Command, Ministry of Defence
	Mr Wolfgang HEUER Desk Officer, Conventional and Humanitarian Arms Control, Ministry of Defence
	Dr Kathrin STEINBRENNER Deputy Head of Division, Federal Foreign Office
India	Mr Siddhartha NATH First Secretary (Disarmament), Permanent Mission of India to the United Nations, Geneva
Israel	Ms Maya YARON Counsellor, Representative to the Conference on Disarmament, Permanent Mission of Israel, Geneva
Japan	Col. Jun KANAI Colonel (Air), First Secretary, Defence Attaché, Permanent Mission of Japan to the Conference on Disarmament, Geneva
Mexico	Lt Col. Omar Leon ARROYO Military Adviser, Permanent Mission of Mexico to the United Nations and other international organizations in Geneva
	Mr Victor M. MARTINEZ ORTA CAMACHO Adviser, Permanent Mission of Mexico to the United Nations and other international organizations in Geneva
Netherlands	Ms Merel EKELHOF * PhD Researcher, Department of Criminal Law and Amsterdam Center for International Law, University of Amsterdam
	Maj. Willem VAN AMERONGEN Senior Legal and Policy Affairs Adviser, Cluster International Affairs, Directorate of Legal Affairs, Ministry of Defence
	Mr Mark VERSTEDEN First Secretary, Permanent Mission of the Netherlands to the Conference on Disarmament, Geneva
Pakistan	Mr Syed Atif RAZA Second Secretary, Permanent Mission of Pakistan to the United Nations, Geneva
Republic of Korea	Mr Jong Heon KIM Military Adviser, Embassy of the Republic of Korea, Bern
	Ms Eun-ji SEO Counsellor, Disarmament, Permanent Mission of the Republic of Korea to the United Nations, Geneva

Russian Federation	Mr Albert EFIMOV * Head of Skolkovo Robotics Centre, Skolkovo Foundation, Moscow
	Mr Andrey MALOV Senior Counsellor, Permanent Mission of the Russian Federation to the Conference on Disarmament, Geneva
South Africa	Mr Cedrick CROWLEY Deputy Director, Conventional Arms, Department of International Relations and Cooperation
	Mr Sipho MASHABA, Deputy Director Compliance, National Arms Control, Department of Defence
	Ms Chantelle NAIDOO First Secretary (Disarmament), Permanent Mission of South Africa, Geneva
Sweden	Mr Martin HAGSTRÖM * Deputy Research Director, Swedish Defence Research Agency, Stockholm
	Mr Daniel NORD Senior Disarmament Officer, Permanent Mission of Sweden to the Conference on Disarmament, Geneva
	Lt Col. Lars OLSSON Adviser, Swedish Armed Forces Headquarters
Switzerland	Dr Vincent CHOFFAT Military Adviser, Permanent Mission of Switzerland to the United Nations and other international organizations in Geneva
	Dr Markus HOEPFLINGER * Research Director – Autonomous Systems, Armasuisse, Federal Department of Defence, Civil Protection and Sport
	Mr Michael SIEGRIST Legal Officer, Federal Department of Foreign Affairs
	Mr Nikolas STÜRCHLER Head of Section, IHL and International Criminal Justice, Directorate of International Law, Federal Department of Foreign Affairs
United Kingdom	Ms Sarah AYLING Conventional Arms Policy Officer, Foreign and Commonwealth Office
	Col. Richard BATTY Lawyer, Development Concepts and Doctrine Centre, Ministry of Defence
	Lt Col. John STROUD-TURP * Conventional Weapons Policy and IHL, Ministry of Defence

United States	Ms Katherine BAKER Policy Adviser, Department of State
	Dr Brian HALL * Interoperability Deputy, Joint Staff Robotic and Autonomous Systems Team, Joint Chiefs of Staff, Department of Defence
	Mr Matthew McCORMACK Attorney, Department of Defence
	Mr Michael MEIER Attorney Adviser, Department of State

INDIVIDUAL EXPERTS (* indicates speaker)

Dr Peter ASARO *	Assistant Professor, School of Media Studies, The New School, US and Vice-Chair, International Committee for Robot Arms Control
Ms Maya BREHM	Researcher, Geneva Academy of International Law and Human Rights, Switzerland
Ms Elena FINKH	Assistant Researcher, United Nations Institute for Disarmament Research (UNIDIR), United Nations Office at Geneva
Prof. Paola GAETA *	Professor and Head of the Department of International Law, The Graduate Institute of International and Development Studies, Geneva, Switzerland
Dr Vadim KOZYULIN *	Senior Researcher, PIR Centre for Policy Studies, Russian Federation
Gp Capt. (Ret'd) Ajay LELE *	Assistant Director, Institute for Defence Studies and Analyses (IDSA), New Delhi, India
Ms Hine-Wai LOOSE	Political Affairs Officer, Implementation Support Unit, Convention on Certain Conventional Weapons (CCW), United Nations Office at Geneva
Mr Richard MOYES *	Managing Partner, Article 36, United Kingdom
Dr Ludovic RIGHETTI *	Group Leader, Autonomous Motion Department, Max Planck Institute for Intelligent Systems, Germany
Dr David RODIN	Co-Director and Senior Research Fellow, Oxford Institute for Ethics, Law and Armed Conflict, University of Oxford, United Kingdom
Ms Heather ROFF *	Senior Research Fellow, Department of Politics and International Relations, University of Oxford, United Kingdom, and Research Scientist, Global Security Initiative, Arizona State University, United States
Lt Col. Alan SCHULLER *	Military Professor, Stockton Center for the Study of International Law, US Naval War College, United States

Dr Gabi SIBONI *	Senior Research Fellow, Director of the Program on Military and Strategic Affairs and the Program on Cyber Security, Institute for National Security Studies (INSS), Israel
Ms Kerstin VIGNARD *	Deputy to the Director and Chief of Operations, United Nations Institute for Disarmament Research (UNIDIR), United Nations Office at Geneva

ICRC

Dr Knut DÖRMANN	Head of the Legal Division and Chief Legal Officer
Ms Kathleen LAWAND	Head of the Arms Unit, Legal Division
Dr Neil DAVISON	Scientific and Policy Adviser, Arms Unit, Legal Division
Dr Gilles GIACCA	Legal Adviser, Arms Unit, Legal Division
Mr Michael RIEPL	Trainee, Legal Division
Mr Mayank VERMA	Trainee, Legal Division

MISSION

The International Committee of the Red Cross (ICRC) is an impartial, neutral and independent organization whose exclusively humanitarian mission is to protect the lives and dignity of victims of armed conflict and other situations of violence and to provide them with assistance. The ICRC also endeavours to prevent suffering by promoting and strengthening humanitarian law and universal humanitarian principles. Established in 1863, the ICRC is at the origin of the Geneva Conventions and the International Red Cross and Red Crescent Movement. It directs and coordinates the international activities conducted by the Movement in armed conflicts and other situations of violence.

ICRC

4283/002 08.2016 500