UNIDIR

Framing Discussions on the Weaponization of Increasingly Autonomous Technologies

No. 1 UNIDIR RESOURCES

Acknowledgements

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities.

In addition, dedicated project funding was received from the governments of the Netherlands and Switzerland.

The Institute would also like to thank Tae Takahashi and Anna Chiapello for their valuable assistance with the expert meeting.

About UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR)—an autonomous institute within the United Nations—conducts research on disarmament and security. UNIDIR is based in Geneva, Switzerland, the centre for bilateral and multilateral disarmament and non-proliferation negotiations, and home of the Conference on Disarmament. The Institute explores current issues pertaining to the variety of existing and future armaments, as well as global diplomacy and local tensions and conflicts. Working with researchers, diplomats, government officials, NGOs and other institutions since 1980, UNIDIR acts as a bridge between the research community and governments. UNIDIR's activities are funded by contributions from governments and donor foundations.

Note

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The views expressed in this publication are the sole responsibility of UNIDIR. They do not necessarily reflect the views or opinions of the United Nations or UNIDIR's sponsors.

www.unidir.org

Framing Discussions on the Weaponization of Increasingly Autonomous Technologies

There are currently discussions in a variety of national and international fora about autonomy and weapon systems. Yet governments are unsure of what they need to know in order to make responsible policy choices—and not all agree that specific policy is necessary. As these are early days in international, multilateral engagement on this issue, this paper seeks to help frame further dialogue on autonomy and weapon systems in a way that is both concise and relevant to policy-making, by helping direct attention to key issues and the areas of greatest concern.

Increasingly autonomous technologies are a feature of today's world—and touch many aspects of our lives—from the factory floor, to the "smart appliances" in our homes, to robot-assisted surgery. Whether we realize it or not, we already rely on machines with considerable autonomy—and this is only going to increase. Advances in robotics, machine learning, artificial intelligence, computational power, networking, engineering and other disciplines are driving increasing autonomy in machines and systems. Such technologies promise (or are already delivering) significant benefits to those who have access to them.

Machines and systems that have increasing amounts of autonomy will require that we carefully examine aspects of our legal codes, privacy regulations, and health and safety policies. These technologies also raise fundamental questions about how we—as societies and individuals—perceive, interact with and relate to machines.

Autonomy is increasingly a characteristic of weapon systems as well. Some consider that autonomy is a desirable characteristic that brings or promises significant strategic advantage and cost benefits. Others have called for legal or policy constraints on particular autonomous functions. Increasing autonomy in weapon systems raises a complicated set of issues, where technological promise, legal regimes, strategic doctrine, moral codes and cultural beliefs about technology are intertwined.

Various policy responses have been put forward including a ban on "killer robots", a moratorium on development of lethal autonomous robots, national self-regulation of development and deployment, and a posture of wait-and-see. However, before embracing one of the available policy responses, many States are still seeking to understand the relevant issues in order to engage constructively in this emerging area.

Given that governments have a responsibility[1] to create or affirm sound policies about which uses of autonomy in weapon systems are legitimate—and that advances in relevant technologies are also creating pressure to do so—UNIDIR launched a project[2] in 2013 that seeks to identify what is important for States to consider when establishing policy relating to the weaponization of increasingly autonomous technologies.[3] This framing paper is based upon a three-day cross-disciplinary expert meeting held in March 2014.[4] The experts represented a range of disciplines—including human rights, international humanitarian law, ethics, artificial intelligence, policy research, defence, military strategy, and evidence-based policy design. Its intent was to bring relevant information into the conversation, build bridges between disciplines and identify areas where different stakeholders—including those in the arms control community, the human rights community, the defence community, the private sector and civil society—might work together on the basis of shared understandings.

Through this project, UNIDIR intends to advance the nascent multilateral discussion by refining the areas of concern, identifying relevant research and linkages, and learning from approaches from other domains that may be of relevance to this topic. Rather than offering specific policy recommendations, the project's primary aim is to provide insights and conceptual frameworks that will enable policy-makers to better think, discuss and make informed decisions about autonomy in weapon systems.

The spectrum of autonomy

Descriptions of autonomy in weapon systems often start with the concept of a technological spectrum moving from remotely controlled systems on one side to autonomous weapon systems on the other. Autonomy increases as one moves along the spectrum from objects controlled by human operators from a distance (such as remotely piloted unmanned aerial vehicles), to automatic and automated systems, to fully autonomous ones.

In modern armed forces there are already a range of systems that can be situated along this spectrum: surveillance devices, range-finding and targeting devices, land-based transport vehicles, aerial vehicles, and robots designed for high threat tasks such as explosive ordnance disposal. Today, these systems are clustered at the lower end of the

1 For States considering introducing new weapons this is a legal responsibility under Article 36 of the 1977 Additional Protocol to the Geneva Conventions, which states: "In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party."

2 For more information about UNIDIR's project "The Weaponization of Increasingly Autonomous Technologies", biographies of the experts involved, and other project materials, see www.unidir.org/programmes/security-and-society/the-weaponization-of-increasingly-autonomous-technologies-implications-for-security-and-arms-control.

3 UNIDIR has purposefully chosen to use the word "technologies" in order to encompass the broadest relevant categorization. In this paper, this categorization includes robots, machines, weapons and weapon systems.

4 UNIDIR would like to acknowledge the thoughtful contributions of the participants of its expert group: John Borrie, Kristian Hammond, Peter Herby, Christof Heyns, Patrick Lin, Noam Lubell, Richard Moyes, Lisa Rudnick, WSP Sidhu, Alexandre Vautravers and Kerstin Vignard. Neil Davison and Eric Steinmyller provided additional expertise. University of Essex Human Rights Centre Clinic students Flavia Colonnese, Ralf Gutmann and Hanna Szabo, supervised by Afonso Seixas-Nunes, provided background research assistance. The views expressed in this paper are the sole responsibility of UNIDIR.

spectrum (remote controlled and automatic/automated). Some States have expressed interest in moving further along this spectrum towards greater autonomy, perhaps as far as "fully autonomous lethal weapons".[5]

An initial hurdle to constructive dialogue on autonomy in weapon systems is that different assessments are made by different States, producers and experts as to where a specific technology sits on the autonomy spectrum. This is compounded by uncertainty surrounding how the object under consideration is labelled: "drones", "robots", "autonomous weapon systems", "killer robots", "lethal autonomous robotics", "lethal and non-lethal" semi- and fully autonomous weapons systems, "supervised autonomy" and other terms. The discussion presently lacks focus, tacking between things (for example, drones, robots and systems), a characteristic (autonomy), and uses (defensive measures? targeting? kill decisions?), in an inconsistent and often confusing way.[6] One of the reasons there are so many different terms being proposed as the object of discussion is that some actors are trying to capture a mix of variables of concern (such as lethality or degree of human control), while others are talking about more general categories of objects.

Some feel that the first step for the international community should be to establish shared definitions of categories or particular thresholds of autonomy along the spectrum described above (using terms such as fully autonomous, semi-autonomous, partially autonomous, supervised autonomy or others). However, this is likely to be a long and complex exercise, compounded by the fact that different incompatible definitions already exist.[7]

There is little ambiguity about the lowest ends of the spectrum: remotely controlled objects are exactly that—humans control the object and make decisions about its actions. However, as one moves along the spectrum its utility as an aid to understanding the problematic diminishes as it doesn't give any indication of the function or functions that autonomy is applied to and how different variables might affect the overall degree of autonomy. The question can be asked "Autonomous in relation to what function"?

There are functions that, when made more autonomous, are considered generally acceptable, such as navigation, transport (such as packbots) and others. Some stakeholders are less at ease with applying increasingly greater autonomy to other functions, such as target selection.[8] There are still other functions that some consider to be of great concern when the characteristic of autonomy is applied—such as the decision to use force and weapons release. Explicitly naming[9] and then creating shared

5 It should be noted that some devices, for example those transporting materiel, are not primarily designed to be lethal. But they may kill or injure though accident or malfunction, or by being equipped with defensive systems.

6 This lack of conceptual clarity is embedded even in the tenses of verbs used in the literature. Present tense is often employed to describe systems that do not yet exist. This creates significant confusion for the reader in trying to determine the actual state of existing technology.

7 For a short overview of the challenges of defining the threshold of autonomy, see Nicholas Marsh, 2014, "Defining the Scope of Autonomy: Issues for the Campaign to Stop Killer Robots", *PRIO Policy Brief no 2*, page 2.

8 So-called "fire and forget" or "launch and leave" missile systems are widely deployed and are sometimes characterized as autonomous. For example, the UK Royal Air Force calls the Brimstone a "fully autonomous, fire-and-forget, anti-armour missile". See United Kingdom Royal Air Force, 2003, *Aircraft & Weapons*, DCC/RAF Publications.

9 Some international organizations, governments and civil society organizations have done exactly that. For example, the term 'critical functions' was used by ICRC in their expert meeting on autonomous weapon systems (26–28 March 2014) to refer to the functions of acquiring, tracking, selecting and attacking targets. Other examples include *Report of the Special Rapporteur on*

understandings among States of which of these functions are of the greatest concern—or raise uncertainties that require further reflection—will help focus multilateral discussions.

Four considerations to help frame discussions

1. Consider the variables that comprise assessments of autonomy

Autonomy is a characteristic of a technology, attached to a function or functions, not an object in itself. A system, for example, might be able to autonomously navigate, yet not be autonomous in selecting its targets. It is a characteristic that might be able to be turned on or off in particular circumstances. There are environments where autonomy is a more beneficial or less risky feature than in others. Not all autonomous functions are of equal concern: some might be uncontroversial while others raise significant legal, ethical, and strategic questions.

When considering autonomy as a characteristic of specific functions in weapon systems, we must also consider that this characteristic is limited or augmented by a variety of other variables. Explicit examination of these variables and how they interact is likely to help States capture the areas of greatest concern and bring focus to the international discussion of policy responses.[10] An initial set of variables might include the following:[11]

- *Goal-satisfying actions*—The ability to create and follow plans of action aimed at satisfying goals. Are goals generated by the system itself or determined by an external source ("orders")? Are plans generated by the system and then vetted through external confirmation (seeking approval) or simply implemented?
- *Predictability*—Predictability of actions the system may take. The simpler the environment, the less the need for a variety of actions and the more predictable a system will become. Likewise, the less variability in the type of actions a system can take, the more predicable it will be, even in a complex environment. The tighter the control that is applied to a system (for example, unwavering focus on a specific goal) the less variable and more predictable a system will be.
- *Communication*—How precise does communication with the system need to be (i.e. does decreasing precision in communication mean the system has to increasingly "interpret" meaning)? How frequent is its communication?
- *Depth of reasoning*—The more limited the reasoning a system is capable of, the less in the way of autonomy it will have and the more predicable it will appear. A simple environment requires less depth of reasoning than a more complex one.
- *Precision of sensors and capacity for synthesis*—The raw sensory capabilities of a system will determine its ability to discriminate things in its environment. The ability

extrajudicial, summary or arbitrary executions, Christof Heyns, UN document A/HRC/23/47 of 9 April 2013; *United States Department of Defence Directive on Autonomy in Weapon Systems*, Directive number 3000.09 of 21 November 2012, and Human Rights Watch and the International Human Rights Clinic at Harvard Law School, 2012, *Losing Humanity: The Case Against Killer Robots*, Human Rights Watch.

10 This is not to suggest extensive discussions of endless permutations of variables, but rather to demonstrate that the term "autonomy" comprises a host of components that need to be explicitly acknowledged as they have direct impact on considerations of both legality and acceptability. By being explicit about these variables we are able to refine the area of concern and set the boundaries of the discussion on the weaponization of increasingly autonomous technologies.

11 Based upon Kris Hammond, 2014, "Autonomous Agents", unpublished.

to combine different types of sensors and synthesize a view of the environment provides a more refined basis for situational awareness.

- *Bounds on location or operating environment*—Control of the physical location and the complexity of the environment in which a system may function will increase control over a system.
- *Functions*—The nature of the actions available to a system (for example navigation, targeting, or weapons release).

The acceptability of any given system will be informed by a careful and well-informed consideration of the interactions of these (and perhaps other) variables.

By way of illustration, the use of highly autonomous systems in remote and uncluttered environments that change slowly, for example underwater systems against sea mines, might have a high level of acceptability.[12] Or consider that highly automated (and some would say autonomous) weapon systems have been deployed for some time, such as C-RAM[13] (Counter Rocket, Artillery and Mortar) systems, anti-ship missile Close-in Weapon Systems (CIWS), and air-defence systems such as Iron Dome. These systems already operate with extremely limited or no real-time human control—they are anti-materiel, defensive in nature, and deployed in uncluttered environments. This combination of variables clearly affects the acceptability assessment. The fact that some of these systems have been deployed for over 30 years while generating little debate attests to that acceptability.[14]

Changing some of the variables listed above to consider the use of highly autonomous weapon systems in complex, rapidly changing or cluttered environments, such as urban areas characterized by a mix of civilians, civilian infrastructure and legitimate military targets, is likely to result in a very different acceptability assessment.

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. If definitions are later determined to be necessary, taking a functional approach now will help inform the definition so that the truly essential elements are captured. It would also reduce the risk that a definition based on technical features is established too early in the process and thus can be easily circumvented later.

2. Consider the drivers

The drivers behind military interest in increasingly autonomous technologies include:

12 The sea environment is particularly attractive for the testing and introduction of increasingly autonomous systems for a range of missions—including counter-mine operations, anti-submarine warfare, and area denial. Experts widely agree that it is likely that we will see a rapid progression from unmanned undersea vehicles to more autonomous ones—as we have seen unmanned aerial vehicles move from being simply remotely operated to systems with increasingly autonomous features.

13 C-RAM systems offer an excellent example of the difficulty of determining unambiguous assessments of thresholds of autonomy—while widely reported to operate in automatic mode, others claim that a human is "in the loop" on every C-RAM engagement. See for example, www.lawfareblog.com/2014/03/guest-post-reflections-on-the-chatham-house-autonomy-conference/.

14 For example, the US Navy has used the Phalanx since 1980.

- greater force projection;
- risk reduction for "dirty and dangerous" missions;
- freeing humans from dull or repetitive tasks;
- dwindling defence budgets coupled with the high costs of military personnel;
- reduced exposure of one's own forces;
- increased speed of decision making, greater accuracy and greater predictability for certain functions in specific environments; and
- a belief by some that autonomous weapon systems may eventually be able to respect international humanitarian law or human rights law better than humans do.

Current military interest centres on increasingly autonomous systems for a limited range of missions, for example force protection, demining, and surveillance of dangerous environments. There is also interest in defensive use to protect borders or military installations, via sensors and/or systems capable of attack. Robotic sentries that have the capacity to be armed are already deployed by countries such as the Republic of Korea and Israel, and other States have announced their intentions to deploy similar systems.[15]

The advantages of physically removing the human from a weapon delivery platform (such as remotely piloted vehicles like drones) is clear as, for example, distance reduces risks to operational personnel, provides for greater endurance than manned platforms, and permits longer periods of observation prior to a decision to use force. However, as autonomy increases, the advantage of removing the human from decision-making is less clear. The claim that superhuman response times will be necessary in future conflicts—and therefore only machines will be able to take such decisions—requires much deeper examination. There might be a difference in acceptability of an autonomous but static system that is a "last line of defence" to counter an incoming attack versus a system that employs superhuman decision-making speed to carry out an attack.

Some military strategists question the necessity or even desirability of delegating responsibility for a decision on launching an attack to autonomous systems. In many militaries, an increasing centralization of military decision-making is occurring, made feasible by modern communications and made necessary due to the political calculations that are essential to ensure mission success and avoid unacceptable political costs for the action taken. Systems that would have the authority to take attack decisions appear to run counter to that trend. From the perspective of centralized command and control, allowing a machine to make decisions related to use of force could appear to be both risky and unnecessary. In addition, there are questions as to whether autonomous weapon systems can contribute to or will undermine the mission of "winning hearts and minds" that characterizes many of today's conflicts.[16]

A second critical yet underappreciated driver is the civilian technology sector—where the robotics revolution is well underway. Indeed this sector is developing a more far ranging

15 For example, Russian Defense Ministry spokesman Maj Dmitry Andreyev recently announced that armed robots would be deployed to guard five nuclear missile launch sites.

16 See, for example, Lieutenant Colonel Douglas A. Pryer, "The Rise of the Machines: Why Increasingly 'Perfect' Weapons Help to Perpetuate our Wars and Endanger Our Nation", paper presented at the 2012 Fort Leavenworth Ethics Symposium, Ft. Leavenworth, Kansas, 3-6 December 2012. Abridged version published in *Military Review*, March-April 2013.

set of autonomous applications than those needed or considered by the military. Industry, the scientific community and even consumers will drive expectations and investment in further advances in autonomous technologies. As civilians become more familiar with increasingly autonomous machines in their daily lives, perceptions of—and trust in—the capacity of machine decision-making will change.

However, the **drivers identified above are not universal and deeper consideration is warranted**. For example, the claim that autonomous systems are more economically sustainable than manned systems is perhaps questionable. Manned skills might shift from, for example, mechanics to software developers, but that doesn't necessarily mean a net reduction in personnel. And, as military forces become increasingly technologically sophisticated, there will be considerable competition from the private sector to employ the highly skilled labour needed to develop and maintain these high-tech systems, potentially driving up costs. While some countries may consider autonomous weapon systems a response to a "manpower crisis", others face no such shortage. This highlights a socio-cultural aspect of the trend towards autonomy that must be taken into account in projecting how military interest in autonomy may evolve globally, regionally and in individual countries.

3. Shift from a technological and innovation-centric frame to one that addresses acceptability, impacts and longer-term consequences of use

As historian David Edgerton reminds us, "the agenda for discussion of the past, present and future of technology is set by the promoters of new technologies": by accepting such an innovation-centric frame, do we risk ignoring equally valid alternatives?

The importance of the introduction of individual new weapons technologies is often overstated.[17] Historically "game changing" technologies often have short-term advantages for those who originally exploit them, but as the technology spreads and is used by more actors and counter-measures are employed this advantage is reduced. Historically we also see that the outcome of most conflicts is decided not by technology but rather by the basic dynamics of the conflict itself, strategy and political interests.

Technology-centric frames are often presented as "all or nothing" in two senses of the term. First, that no other alternative exists to a particular innovation, and second, that there is only a choice between a world without the technology and a world where the technology is fully integrated.

This is misleading. First, "old" technologies will be deployed and used alongside new ones. This fact raises a complex set of challenges—compatibility between old and new systems, developing "make do" patches to facilitate interactions between them, untested interactions between old and new systems, etc. The second reason this is misleading is perhaps more important: such a frame can artificially limit our choices to choosing between technology or forgoing it altogether.[18] What about a hybrid approach that plays to the strengths of

17 For a thought-provoking analysis of how innovation-centric perspectives (as opposed to use-centric ones) distort how we view specific technologies and their historic importance, see David Edgerton, *The Shock of the Old: Technology and Global History since 1900*, Profile Books, 2006. For examples of this in relation to specific weapons technologies, see disarmamentinsight.blogspot.ch/search?q=edgerton

18 This is not unique to the consideration of autonomy in weapon systems. For example, the Nuffield Council on Bioethics highlighted this in its thorough consideration of another emerging area—biotechnology. They stated that "… commitments to particular technological pathways should be evaluated not only in terms of their expected future impacts but also by comparison to possible

both man and machine?[19] Further exploration is needed on how to harness the benefits of certain uses of autonomy without sacrificing our humanity or the humanity of others.

Humans have a poor track-record of predicting the full range of benefits or risks associated with new technologies. Often technologies are developed for one set of tasks but then adopted in other fields for missions not envisaged by the designers or proponents. Concerns have been raised, for example, that deployment of increasingly autonomous weapon technologies will begin in uncluttered environments and steadily migrate into more complex ones, perhaps without a State undertaking a new Article 36 review (see footnote 1) or policy review that takes into consideration the different operating environment. In addition, humans, including armed forces in battle environments, have a tendency to manipulate and modify technologies to overcome safety features and controls.

Peter Singer, author of *Wired for War*, has noted the danger of technology-centric frames: "… too often in discussions of technology we focus on the widget. We focus on how it works and its direct and obvious uses. … Indeed, with robotics, **the issues on the technical side may ultimately be much easier to resolve than dilemmas that emerge from our human use of them**."[20]

There is a rich and growing literature on the legal, policy and ethical dimensions of the weaponization of increasingly autonomous technologies. While this work continues, increasing attention to broader risks and potential long-term implications would be useful. There is a lack of critical analysis on, for example: how the proliferation of increasingly autonomous systems might alter regional security dynamics; whether increasingly autonomous weapon systems will drive development of other weapons of concern, counter-measures or methods including cyber-conflict; and whether the weaponization of autonomous technologies will increase asymmetric warfare and terrorism, to name a few. Increasing interaction with other stakeholders—including the private sector, research, and civil society, as well as from a greater variety of countries—will help policy-makers have a more nuanced understanding of the issue. In all areas the perspectives of the developing world need to be solicited and heard.

A significant amount of what is known about military interest in autonomy comes from only a few countries. Other States' perceptions of utility, necessity, desirability and consequences are critical to having a well-informed and balanced conversation. The Human Rights Council, meetings on this subject within the framework of the Convention on Certain Conventional Weapons (CCW) and other fora will offer States an opportunity to voice their concerns and convictions, thereby broadening and deepening the discussion.

4. Beware of how you define the area of concern

The widely used shortcuts adopted to make sense of the complexities of the discussion (such as the man "in", "on" or "out" of the loop description, or the categories of remotely

---

alternative pathways; this can help to illuminate obscured assumptions, constraints and mechanisms of the innovation system, and help to identify sites and opportunities for more constructive governance, prioritisation and control." See Nuffield Council on Bioethics, 2012, *Emerging Biotechnologies: Technology, Choice and the Public Good*, para 10.5, p. 175.

19 For example, see Noel Sharkey, 2014, Towards a principle for the human supervisory control of robot weapons, *Politica & Società*, no. 2, May-August.

20 P. Singer, *The Robotics Revolution*, Brookings Institution, 11 December 2012, www.brookings.edu/research/opinions/2012/12/11-robotics-military-singer

controlled, automatic, automated and autonomous technologies), are perhaps not the only, or even the most useful, distinctions to make. If adopted without careful thought to broader or longer-term issues, such approaches may draw attention away from others that could be more productive in characterizing and responding to the challenges associated with the weaponization of increasing autonomous technologies.

Is the concept of "lethality" helpful or limiting?

- Is there a concern about weaponization of increasingly autonomous technologies that target purely materiel military objectives?
- Is it also an issue that machines might one day take autonomous decisions to deploy so-called non-lethal force, for example firing rubber bullets, riot control agents or beanbag rounds?
- Would other distinctions, such as "the use of force" or the broader terminology of "use of weapons in attacks", taken from international humanitarian law, be more useful than a focus on "lethal"?

Increasingly autonomous technologies are unlikely to be limited to contexts of armed conflict. If widely deployed, it is likely that they will also be employed in domestic law enforcement and even in peacekeeping operations. Thus, national domestic consideration of the acceptability of increasingly autonomous weapons is also necessary. While international fora, such as the CCW, have mandates that restrict wider consideration of the issue, national cross-ministry discussions should take into consideration potential domestic use as well as military applications in order to ensure coordinated and coherent national policy.

Man in, on or out of the loop?

- Does the categorisation of a human in, on or out of the decision-making loop distract from other, perhaps more useful, distinctions, such as the type of decisions, their quality, and the responsibility for decisions taken?

Discussions often assume that having a human "in" or "on" the loop ensures rigorous human decision-making concerning the launching of an attack. Already today, a human decision to "fire" or "attack" with a given weapon system may depend on large amounts of data and analysis—collected, processed and interpreted by both humans and machines. As weapon systems are granted increasing amounts of autonomy, how useful is this "loop" distinction if there isn't a corresponding discussion of what is the human's role, value and responsibility in a decision-making process?[21]

Meaningful human control?

- What is the nature of human control to be exercised over increasingly autonomous systems? [22]

21 As noted by the NGO Article 36, questions about the significance or value added of having a human in or on the loop might arise, for example, "if that person simply pressed a 'fire button' every time a light came on without having any other information." See Article 36, 2013, *Structuring Debate on Autonomous Weapons Systems*, www.article36.org/wp-content/uploads/2013/11/Autonomous-weapons-memo-for-CCW.pdf.

22 This question and others are raised by Article 36. For an introduction to the concept of meaningful human control and key questions that arise from its consideration, see the memorandum for CCW delegates written by Article 36, *Structuring Debate on Autonomous Weapons Systems*, ibid.

- Are the controls that are "programmed" into a system and that set the bounds of the mission (such as the geographical area of operation, the time dimension, what to do in case of an ambiguous situation)[23]—an adequate form of human control or a replacement for real-time control?
- In the operation of increasing autonomous systems, what constitutes meaningful human control over specific individual attacks?

The concept of meaningful human control may be a useful approach to engaging in discussions aimed at analysing and/or regulating the weaponization of increasingly autonomous technologies. One way to help focus such a discussion would be to address "meaningful human control" of "attacks", which are a defined concept in international humanitarian law.[24]

An approach based on meaningful human control has the advantage of focusing on the responsibilities of humans, rather than focusing on continuously evolving technological capabilities. It might provide common ground for evaluating specific technologies, in light of how meaningful human control is to be exercised. It also provides an entry point to the discussion that is accessible to a much wider range of actors including those with limited technical expertise.

Looking forward

Emerging technologies raise novel and challenging legal, ethical and strategic issues. Consideration of these issues by policy-makers entails the additional challenge of their gaining familiarity with the scientific or technical concepts, innovations, techniques and materials involved. Technologies are likely to evolve faster than decision makers can learn about them. As the pace of development is likely to only increase, perhaps it would be most useful to direct attention in policy discussions to **consideration of the variables that factor into acceptability assessments, as well as drivers, risks and the longer term consequences** of the characteristic of autonomy in weapon systems, and away from defining technological thresholds, timeframes and technological fixes.

There is considerable disagreement among experts on the state of the development of component technologies—particularly in artificial intelligence and machine learning—and the timelines for their integration in weapon systems. That said, robotic autonomy writ large is an area of extremely high investment by the private sector as well as the military and thus **the components of increasingly autonomous technologies (for both civilian and military applications) will continue to improve, even if the pace of improvement is open to question**.

Therefore, **it is not too early to engage in discussions on potential consequences and ethical considerations**. It is crucial to give significant consideration today to the question "If a weapon system were ABLE to do X, would we WANT it to do so?" This question offers an opportunity for reflection that goes beyond IHL-based legal assessments to other fundamental considerations such as the right to life and protection of human dignity. This discussion will benefit from increasing interactions between the security community with that of human rights, defence, and other disciplines more widely, including ethics and philosophy, science and technology, psychology, sociology, engineering, and others.

23 Such as the variables noted on page 4.

24 For example, Article 49(1) of the 1977 Additional Protocol I to the 1949 Geneva Conventions defines "attacks" as "acts of violence against the adversary, whether in offence or in defence."

In addition, it is important to acknowledge that the discussion of increasing autonomy in weapon systems is serving as a proxy for a variety of other issues—from fundamental questions about humans' relationship with technology and visions of the future, to discomfort with how armed unmanned aerial vehicles have been used in countries such as Pakistan and Yemen, to questions about the concepts of remoteness, risks and fairness in conflict. It is notable that the issue of autonomy has opened the doors for discussion on a broad range of worthy issues that until now have not coalesced around a single topic. At the same time it is worthwhile to consider **how many of these issues are unique to the topic of the weaponization of increasingly autonomous technologies**. Once examined, it might emerge that the issue of greatest concern isn't necessarily the technological one—but rather the appropriate, legal and responsible role of humans—today and in the future—in the decision to employ force.

The technological capability for autonomous weapon systems that can detect and analyse complex environments, select targets and carry out an attack is likely to be reality one day—even if that day is far in the future. But **the decision to weaponize these capabilities is not inevitable**. International and national discussions must centre around which applications of these capabilities are acceptable, legal, and desirable when applied to the use of force.

What are the benefits, challenges and risks represented by the weaponization of increasingly autonomous technologies? Likewise what might be the long-term consequences—intentional and unintentional—of developing and deploying increasingly autonomous weapon systems? Broad and well-informed discussions of these topics will help us identify the areas that evoke clearer responses that can be the basis of policy development, and those that entail greater ambiguity or uncertainty—and therefore require deeper consideration and further analysis by States, researchers, the private sector and civil society.

UNIDIR

Framing Discussions on the Weaponization of Increasingly Autonomous Technologies

There are currently discussions in a variety of national and international fora about autonomy and weapon systems. Yet governments are unsure of what they need to know in order to make responsible policy choices—and not all agree that specific policy is necessary. As these are early days in international, multilateral engagement on this issue, this paper seeks to help frame further dialogue on autonomy and weapon systems in a way that is both concise and relevant to policy-making, by helping direct attention to key issues and the areas of greatest concern.