UNIDIR

The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches

a primer

No. 6

UNIDIR RESOURCES

Acknowledgements

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. In addition, dedicated funding for this project was received from the Government of Germany.

About the Project "The Weaponization of Increasingly Autonomous Technologies"

Given that governments have a responsibility 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's work in this area is focused on what is important for States to consider when establishing policy relating to the weaponization of increasingly autonomous technologies. See http://bit.ly/UNIDIR_Autonomy for Observation Papers, audio files from public events, and other materials.

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

Contents

Acronyms and abbreviations

AGI	Artificial General Intelligence
AI	Artificial Intelligence
AWS	Autonomous Weapons Systems
CCW	Convention on Certain Conventional Weapons
CID	Combat Identification
C-RAM	Counter Rocket, Artillery and Mortar
DoD	Department of Defense
GGE	Group of Governmental Experts
GOIS	Going onto an Object in Space
GOT	Going onto a Target
ICRC	International Committee of the Red Cross
IHL	International Humanitarian Law
LAWS	Lethal Autonomous Weapons Systems
L-RASM	Long Range Anti-Ship Missile
SoS	Systems of Systems

Introduction

In 2016, the High Contracting Parties to the Convention on Certain Conventional Weapons (CCW) mandated the establishment of a Group of Governmental Experts (GGE) to “explore and agree on possible recommendations on options related to emerging technologies in the area of [Lethal Autonomous Weapon Systems or LAWS],”[1] and they noted that they should consider “identification of characteristics and elaboration of a working definition of LAWS”. This primer is in support of that endeavour.

The High Contracting Parties have tasked themselves with identifying relevant characteristics, elaborating a working definition, and ultimately agreeing on recommendations. Since 2013, in the CCW informal meetings of experts, concerns, characteristics and definitions have been discussed concurrently rather than sequentially.

Some concerns, such as ethical and legal ones, have been present since the beginning of the international discussion. Others, such as issues of risk, safety and bias, have emerged as the conversation has deepened and become more nuanced.

At the heart of much of the CCW discussion has been identification and discussion of a multitude of desirable or undesirable characteristics—such as practicability, mobility or accountability—potentially related to or describing features of autonomous weapon systems. Many of these terms have multiple meanings and it has not been always evident which meaning was intended, and sometimes they have been used interchangeably. Greater conceptual clarity about these terms will help focus the work of the GGE.

Agreeing on a working definition of LAWS will be a challenging endeavour, as there are several definitions already in circulation, and some stakeholders have already stated a preferred policy response. Moreover, each proposed definition attends to a particular set of concerns and characteristics, while omitting others.

One’s position on both an appropriate definition and an adequate policy response ultimately depends on what one is concerned about. Different definitions will attend to different sets of concerns, as well as privilege different sets of characteristics.

The objective of this primer is to consolidate and give an overview of both concerns and characteristics and illustrate how different definitional approaches attend to these.

This paper has five sections:

1. A brief mapping of concerns that have been raised in the international discussion;
2. An exploration of some of the characteristics that have been raised in relation to LAWS, yet are often understood to mean different things;
3. A description of different definitional approaches;
4. A selection of proposed definitions and how they attend to different concerns and characteristics; and
5. Conclusions.

This primer is in no way exhaustive. It is rather an attempt to support High Contracting Parties as they determine a clear and logical approach to the GGE’s discussions.

[1] CCW/V/V2, para 3. Lethal Autonomous Weapons Systems or “LAWS”. Additionally, this paper will refer to autonomous weapon systems (AWS), but where appropriate for citation or clarification use LAWS to only refer to those set of systems that possess lethality.

1. Concerns

The earliest days of the international discussion on autonomous weapon systems focused on human rights and legal concerns, but over time, as governments have developed a deeper appreciation of the issues surrounding autonomy in weapon systems, additional concerns have been brought to the table.

This section contains a brief description of the spectrum of concerns that have been articulated by governments and other stakeholders.

Human rights and ethics

The human rights and ethical dimension of the Autonomous Weapons Systems (AWS) question was the first concern to attract international attention. In the spring of 2013, the UN Human Rights Special Rapporteur on extrajudicial, summary, or arbitrary executions, Christof Heyns, released a report[2] in which he recommended a moratorium on the development of what he called "Lethal Autonomous Robots". He called for this moratorium on the grounds that these so-called "killer robots" might pose significant challenges to the right to life and the right to human dignity.

Since then, the ethical questions have generated much discussion.[3] Questions range from whether decisions to intentionally take a life ought to be delegated to an object, to others arguing that there is a moral responsibility to develop and deploy autonomous weapon systems if, through greater precision and situational awareness, they lower civilian harm or increase protection of one's own forces.

To date, the international policy discourse between States has privileged concerns about the law of armed conflict over situations outside armed conflict in which human rights law pertains, or to broader ethical concerns. Much of the ethical discussion that has occurred is grounded on interpretations of the Martens Clause[4], by reminding States that even in situations where there is no specific law, combatants and non-combatants remain under the protection of the principles of humanity and the dictates of the public conscience. In this way, the law draws off of the normative power of ethical principles to regulate the development or use of weapon systems in international law.

However, while rights and ethical concerns have been prominent in the civil society discourse on AWS, the CCW discussion has spent more time on issues concerning compliance with international humanitarian law (IHL) and technical considerations.

[2] See C. Heyns, 2013, *Report of the Special Rapporteur on extrajudicial summary or arbitrary executions*, United Nations Human Rights Council, A/HRC/23/47, www.ohchr.org/Documents/HRBodies/HRCouncil/ RegularSession/Session23/A-HRC-23-47_en.pdf.

[3] For a detailed discussion of ethics and values, see UNIDIR, 2015, *The Weaponization of Increasingly Autonomous Technologies: Considering Ethics and Social Values,* UNIDIR Resources no. 3, http://www.unidir.org/files/publications/pdfs/considering-ethics-and-social-values-en-624.pdf.

[4] The Martens Clause, originally found in the preamble to the 1899 Second Hague Convention and most recently set out in 1977 Additional Protocol II, states that "in cases not covered by the law in force, the human person remains under the protection of the principles of humanity and the dictates of the public conscience."

Legality[5]

The CCW is an international arms control treaty whose purpose is to ban or restrict the use of specific types of weapons that are considered to cause "unnecessary, unjustifiable or superfluous suffering to combatants or to affect civilians indiscriminately."[6] Its modular and flexible nature—with a short Convention and then protocols negotiated as needed and attached to it—seems well adapted to respond to new developments in weapons.

By situating the discussion on autonomous weapons within the CCW, the international community has framed the discussion as primarily a humanitarian law one: whether increasingly autonomous weapons will be able to comply with the laws of war, and specifically the IHL principles of necessity, proportionality, and distinction. It has also engendered a discussion of how autonomous weapons that could have learning algorithms embedded in them can be tested in a way that respects States' commitments to review new weapons, methods and means of warfare under Article 36 of the 1977 Additional Protocol I to the Geneva Conventions. Questions of legal accountability for the use of autonomous weapon systems, as well as the roles and responsibilities of humans in decisions to use force, have also been raised (see Section II).

In this framing, the technological barrier (see below) to the development of autonomous weapons is linked to the legal barrier—once the technology is proven to be able to meet the legal standards, there is no existing legal impediment to their development, deployment or use as long as they continue to meet these standards.

Technological[7]

As so much of the AWS discussion is speculative, and there are widely different assessments by technologists and governments alike on the speed and trajectory of relevant developments, it has been challenging to anchor the policy discussion in purely technical assessments. Technological concerns range from how to ensure predictability and reliability in increasingly autonomous systems, to how to reduce risks of unintended interactions, to how can governments design, test, and verify their autonomous systems. This technological discussion extends to whether existing mechanisms, such as Article 36 weapon reviews, are adequate or appropriate for regulating AWS.

In September 2017, in preparation for the November GGE meeting, the Chairman produced a "food-for-thought" paper focusing on the current state of relevant technological developments and their incorporation in specific military systems.[8] In this paper, the relevance of Artificial Intelligence (AI) is explicitly brought to the fore, as well as the issues of verifiability and scrutability of autonomous systems.

[5] A wide variety of views on legal interpretations area available. See, for example, International Committee of the Red Cross, 2016, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons,* pp. 79–82, https://www.icrc.org/en/publication/4283-autonomous-weapons-systems#;
Human Rights Watch and International Human Rights Clinic, 2015, *Mind the Gap: The Lack of Accountability for Killer Robots*, Human Rights Watch, https://www.hrw.org/sites/default/files/reports/arms0415_ForUpload_0.pdf; and Kenneth Anderson and Matthew Waxman, 2013*, Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can,* Hoover Institution, http://media.hoover.org/sites/default/files/documents/Anderson-Waxman_LawAndEthics_r2_FINAL.pdf.

[6] See https://www.unog.ch/ccw.

[7] Paul Scharre and Michael C. Horowitz, 2015, *An Introduction to Autonomy in Weapon Systems*, Center for a New American Security; and Vincent Boulanin, 2016, *Mapping The Development of Autonomy in Weapon Systems: A Primer on Autonomy,* SIPRI, https://www.sipri.org/sites/default/files/Mapping-development-autonomy-in-weapon-systems.pdf.

[8] Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, *Food-for-Thought Paper Submitted by the Chairperson (advance version),* 4 September 2017.

Proliferation and arms-racing

Over the past three years of CCW informal meetings of experts, many stakeholders noted the danger of AWS proliferation and have suggested that a regulatory response now would be the best hedge against a future proliferation problem. However, even if such systems are regulated in some manner, it will remain difficult to fully restrict their proliferation. With smaller and more efficient systems—with readily available dual-use components—there lies an ability for such systems to proliferate to individual actors or groups with malicious intent, such as terrorists.

Additionally, with many militaries currently undergoing significant modernization efforts, there is an incentive to constantly pursue increasing autonomy to maintain a technological advantage, with adversaries responding in kind and creating a traditional arms-race dynamic.

Furthermore, as some militaries begin to develop autonomous systems, others will look to drive development and use of countermeasures for such weapons. In particular, there will be incentives to use cyber and information operations to penetrate not only the physical weapons systems, but command, control and communications networks that provide autonomous systems with information. We are also likely to see a growing interest in cyber operations and increasingly powerful electromagnetic weapons. UNIDIR has raised the issue of AWS vulnerabilities to cyber operations for some time,[9] and notes that the 2017 Chairman's food-for-thought paper asks whether AWS would be susceptible to hacking.

Strategic stability[10]

As States rush to lead or dominate the development of AI, they may become more accepting of risk in applying those developments to weapons. As Russian President Putin recently stated: "artificial intelligence is the future, not only for Russia, but for all humankind," and "it comes with colossal opportunities but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become ruler of the world."[11] Such attitudes create incentives to be a "first mover" in the field of AI, but also in its applications.

Given the incentives to be a first mover, and the conditions for arms race dynamics as described above, increasingly autonomous weapons might create regional or global instability or lower the threshold for the use of force. As belligerents may be less concerned with force protection, States may extend the use of autonomous weapons to strategically sensitive tasks or roles. For instance, a State may use increasing autonomy to secure its territory, protect its borders or strategic assets, or engage in counter-terrorism operations. Additionally, with increasing autonomy distributed in a battlespace, there will be incentives to shorten time cycles between decision and action. This potential for a "flash war" may be highly destabilizing.

[9] UNIDIR hosted an expert group meeting on the intersection of autonomy, AI and cyber operations in November 2015, has highlighted this issue in its annual statements in CCW, has held events such as "Cyber Weapons and Autonomous Weapons: Potential Overlap, Interaction and Vulnerabilities" (9 October 2015, listen to the presentations at http://unidir.org/programmes/emerging-security-issues/the-weaponization-of-increasingly-autonomous-technologies-phase-iii/cyber-weapons-and-autonomous-weapons-potential-overlap-interaction-and-vulnerabilities); and has issued an Observation Report on this topic.

[10] For more on strategic stability issues, listen to the presentation by Paul Scharre at the April 2016 UNIDIR event "Understanding Different Kinds of Risk", http://www.unidir.org/programmes/emerging-security-issues/the-weaponization-of-increasingly-autonomous-technologies-phase-iii/understanding-different-types-of-risks.

[11] https://www.rt.com/news/401731-ai-rule-world-putin/; see also Juergen Altmann and Frank Sauer, 2016, "Speed Kills! Why we need to hit the brakes on 'killer robots'", ICRAC, https://icrac.net/2016/04/speed-kills-why-we-need-to-hit-the-brakes-on-killer-robots/.

The dual-use nature of the technologies

Managing the international security aspects of dual-use technologies is not new. Indeed, the international community has regulated dual-use technologies and materials in the past through mechanisms such as the Chemical Weapons Convention and the Biological and Toxin Weapon Convention. The arms control community also has significant experience with technology control regimes—for example the Missile Technology Control Regime, the Nuclear Suppliers Group, the Wassenaar Arrangement, etc., where a group of States control access to particular items or materials for reasons of international security or non-proliferation concerns.

Today, the spread and potential misuse of many ubiquitous emerging technologies have international security implications—yet these technologies have a much wider group of stakeholders—both respectable and unsavoury—and a hugely vested private sector. Some governments have already articulated their concern that regulation of AWS means that they will be denied technologies and locked out of extremely important high-tech sectors, or that development of civilian applications of increasing autonomy will be harmed. Together these factors make traditional responses, such as control regimes, less likely to succeed.

The challenge of verification, which has received insufficient attention thus far in the AWS discussion, could be informed by examination of other regimes, such as the BTWC. Lessons might be drawn, for example, from the unsuccessful efforts to negotiate a mechanism to verify the BTWC, which broke down in the face of the challenges of regulating dual-use materials and the reluctance of key States to restrain their industries or risk loss of proprietary information.

Operational concerns

There are several pressing operational concerns regarding AWS. As there is uncertainty whether such systems will work as intended, as well as regarding their reliability and predictability, there is worry whether commanders will be able to trust the systems and thus use them during armed conflict. They may be unwilling to use a system that would hold them accountable, or strictly liable, for the unforeseen or unintended acts of it. Where emergent behaviour is more likely to occur, such as in swarms of AWS, there may be operational concerns that the commander would be unable to control the systems appropriately. In joint operations, moreover, there are concerns about the interoperability of different systems acting either independently or collaboratively.

Furthermore, research into human factors analysis and machine interfaces suggest that humans are likely to either over- or under-trust the performance of certain systems. For example, if one were to have AWS embedded within units of soldiers, there may be over-confidence in the abilities of the systems, thereby leading to dangerous or risky uses, or their use in situations where AWS fail to provide the presumed protection to soldiers.[12]

Finally, there is simultaneous hope and concern that AWS are required to support new missions previously viewed as too costly to pursue, or previously not possible.

[12] This has been a long-identified issue. See, for example, Missy Cummings, *Automation Bias in Intelligent Time Critical Decision Support Systems,* American Institute of Aeronautics and Astronautics 1st Intelligent Systems Technical Conference, 2004, p. 1, p. 5, wayback.archive.org/web/20141101113133/http://web.mit.edu/aeroastro/labs/halab/papers/CummingsAIAAbias.pdf.

Unintentional risk and safety issues[13]

As noted in a previous UNIDIR report:

> Any complex, hazardous technology carries "unintentional" risk, and can have harmful results its designers and operators did not intend. AWS may pose novel, unintended forms of hazard to human life that typical approaches to ensuring responsibility do not effectively manage because these systems may behave in unpredictable ways that are difficult to prevent."[14] "History shows that certain kinds of system for the management of hazardous technology possessing significant levels of automation can fail in ways not anticipated by their human designers and operators. … [C]atastrophic failures occur despite careful technological design and planning, organizational control and training, and the addition of multiple technical redundancies.[15]

Autonomy "at rest" or "in motion"[16]

For the most part, discussions at CCW have focused on physical "in motion" weapon systems; that is, systems that are able to act on and in their environment. Yet an important military application of autonomy is in "at rest" systems, such as decision aids. These systems are not directly coupled to a munition, yet are used in support of decisions to use force, such as selecting target sets, conducting proportionality calculations, and war-gaming potential courses of action. As stated in a previous UNIDIR report:[17]

> Systems that process large amounts of sensory and intelligence data in order to aid military decision making and logistics planning hold obvious appeal for the military advantage this might convey[18]… . This has major implications for safety because, for all of their promise, machine learning-based systems present challenges:
>
> - Machine learning systems, and in particular neural networks and similar architectures, are **complex**. Their effectiveness is the result of their mathematical properties and complex relationships between opaque internal parameters. It means that even the researchers running the systems do not have a complete understanding of the underlying learned logic of, say, a trained deep learning network.
> - Currently it is not possible to produce **formal proofs** of the behaviour of machine learning systems. This poses challenges for attaining the levels of formal verification that are demanded for many software code-based systems, especially for systems performing critical functions on which human lives may rely.

[13] See also Paul Scharre, 2016, *Autonomous Weapons and Operational Risk,* Center for a New American Security, https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf?mtime=20160906080515.

[14] UNIDIR, 2016, *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies,* UNIDIR Resources no. 5, p. 1, http://www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf.

[15] Ibid, p. 6.

[16] See United States Department of Defense, 2016, *Defense Science Board Summer Study on Autonomy*, especially chapter 4 ("Strengthening operational pull for autonomy"). The CCW Chairperson's food-for-thought paper characterizes this as discrete systems or spread-out information processing systems.

[17] UNIDIR, 2016, *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies,* UNIDIR Resources no. 5, p. 8, http://www.unidir.org/files/publications/pdfs/safety-unintentional-risk-and-accidents-en-668.pdf.

[18] See United States Department of Defense, 2016, *Defense Science Board Summer Study on Autonomy*, ch. 4.

- Machine learning systems are stochastic, and **so predictability is a challenge**. The machine is not constrained by human experience or expectations.[19] Systems can be tested, and their behaviour observed in a range of scenarios—but this is a long way from formal verification.
- **Interpretability** (the ability to analyse and assess the "learnt logic" on a machine learning system) is in its infancy. At present, techniques are crude.
- Machine learning systems tend to be **tightly coupled**. Many applications involve a single deep learning network that is largely opaque once it is trained.

Immediacy of the concern

There are divergent views as to whether AWS should be characterized as a near-term, mid-term or long-term concern. This is very much linked to what one imagines an autonomous weapon to be—simply "smarter" versions of existing systems or actual intelligent robotic soldiers. In technological terms, this could be summarized as whether the concerns are about narrow applications of Artificial Intelligence (AI) in weapons systems or Artificial General Intelligence (AGI).

Influencing this assessment should be the recognition that constant innovations in hardware and software, particularly advances in AI and the power to compute, mean that weapon systems are becoming increasingly autonomous in an incremental fashion. For example, new guidance systems enable systems to be without communications for longer periods, and advances in power sources and flight physics permit ever-longer loitering times. These incremental improvements are occurring at a more rapid rate in a wider variety of areas and are, for the most part, arising from the private sector. Advances in greater autonomy, therefore, is unlikely to be steady and linear, but more likely to be rapid, intermittent and nonlinear—and developments will not necessarily be under the control of militaries.

When we consider the spectrum of increasingly autonomous weapon systems, it is easier to conceptualize what autonomy looks like at the far end of the spectrum—the Terminator-type scenarios. Defining and regulating that end of the spectrum might be easier than imagining the incrementally increasing autonomy from where we are today. At what point does it "tip" from just being a gradual improvement in current systems to being an object of concern? Deciding if we are concerned about nearer-term potential developments in weapon systems or only far future ones (regardless of the differences in opinion of how far off the far end of the spectrum is) will ultimately determine the urgency of the policy response.

****

This section provided an overview of the variety of diverse concerns that stakeholders have articulated about increasing autonomy in weapon systems. Ultimately, the CCW might not be the appropriate forum to address all concerns, due to its format, membership, resident expertise, focus, pace or other reasons. It is clear that there are strongly held beliefs about all of these concerns; therefore, it would be productive for States to consider whether and how each concern could be addressed within the CCW, and if not, what the appropriate forum to do so might be to ensure that the concerns that cannot or will not be resolved in the CCW do not inhibit or impede progress in areas where the CCW can advance.

[19] For a discussion, see J. Tapson, 2016, "Google just proved how unpredictable artificial intelligence can be", *Business Insider UK*, 19 March, uk.businessinsider.com/google-just-proved-how-unpredictable-artificial-intelligence-can-be-2016-3.

2. Characteristics

A variety of characteristics and features have been identified as relevant to the AWS discussion. However, these terms are often used to mean different things or are used in an imprecise way. As governments start to turn their attention to the question of definitions, it is worthwhile to briefly consider these features in turn.

Automation or autonomy

In international discussions, there is a wide spectrum of views of what “autonomy” means. Some claim, for example that a landmine is an autonomous weapon since there is no “human control” over when it detonates. On the other end of the spectrum, some proposed definitions of an AWS describe objects that are so independent, they stretch belief that any responsible State would be interested in adding such a weapon to its arsenal.

Automated weapon systems are nothing new. Crudely put, if one has a pre-programmed system—whether in code or mechanically—with "if this, do that" logic, then one is talking about an automated weapon system. The system is simply carrying out the task and it has all the human-embedded answers ahead of time. On the mechanical end of the spectrum we have weapons like landmines, where the object is engineered to detonate when particular conditions are met—often a pressure plate triggered at a specific weight load. A landmine doesn’t “decide” whether or not to detonate.

But as one moves away from the automation side of the autonomy spectrum, it becomes more difficult to draw firm lines. For example, Phalanx CIWS, a defensive weapon system against anti-ship missiles, has been deployed for over 30 years. Its land-based equivalent, C-RAM (Counter Rocket, Artillery and Mortar), is used to detect and destroy incoming artillery, rockets and mortar rounds. The system is programmed to recognize particular rocket and mortar signatures. While these systems have a human-supervised mode, once the system is “on”, it will destroy incoming objects that match the signature set without further human engagement based on its faster-than-human response times.

The international discussion on AWS has not been about these sorts of existing, already long-deployed systems. However, if these systems are already operating without further or real-time human engagement, what is the “incremental increase” in autonomy that has raised the concern of some in the international community? Is it the potential for moving from anti-materiel to anti-personnel systems? Is it moving from defensive applications to offensive ones? Is it the capability of a weapon system to undertake an action at the time and place of its choosing, without that decision being vetted or approved in real time by a human being? Is it that progressively more sophisticated applications of AI promise the ability to model and predict future actions with increasing accuracy, and therefore encourage more aggressive, first-strike postures? Or that increasingly sophisticated and opaque AI means that we will understand less and less about how decisions or recommendations are made? Rather than pointing to a bright line between what might be considered autonomous or not, these questions indicate that we need deeper discussions to home in on what specific aspects, characteristics, or applications of autonomy require the international community’s attention.

Not all applications of autonomy are created equal. An important step in making progress on refining the area of concern will need to be explicit and specific about where autonomy is applied in a system. In the discussion on AWS, it is unlikely that the true concern is whether a weapon system can, for example, navigate autonomously. But States might care greatly about how much autonomy a system has once it gets in vicinity of the target, to select among targets, or when and with which means to engage them. Autonomy applied to tasks other than the so-called “critical functions” of selecting and engaging targets are ultimately unlikely to be a serious concern.

Learning, adaptation and adjustment

Some machine learning systems are able to learn through simulation or direct experience or a combination of both. This *learning* can be supervised; that is, with humans labelling all of the training data and correcting errors. Or, learning could be unsupervised, where the system learns the underlying structure of the data itself without it being labelled. Learning can also be done “offline” where the system learns its task by being provided a static dataset. No new data comes into the system. It can then be “frozen” after it reaches a particular threshold set by its creators and cannot continue to learn while in use. Other systems can continue to learn, through “online” learning, where the data inputs constantly change, and thus continue to update its model of the world and its parameters. These continuous learning systems are said to “adapt”. *Adaptation* is the ability to change with environmental inputs. These systems continually update their internal states and representations, as well as their decisions, based on external stimuli and the probability distribution of those stimuli.

From a technical perspective, any system that continues to learn while deployed is constantly changing. It is not the same system it was when deployed or verified for deployment. Some have raised questions about the legality of adaptive systems, particularly in regards to States’ Article 36 obligations.[20] One may test, evaluate, validate and verify a system at one point in time, but very quickly the system changes so that it is no longer the same system with the same data inputs or parameters.

Adaptation is not a feature of “automated” systems. They will act, mechanistically, regardless of environment and only upon receiving a particular input (and maybe at a particular time) and will act in one particular way (the output). However, with adaptation, or the possibility of adaptation, we create the potential for different possibilities.

Adjustment is also different from adaptation. In a system that possesses adjustment, there are various “modes” that one might be able to “dial” up or down. In these instances, there is typically a human operator making the decision about which mode is appropriate to the situation. For example, the Patriot Air and Missile Defense system has several modes—it can be operated manually, semi-automatically, or be “fully automatic”. The automaticity does not change, just the range of actions, speed, or human-machine interaction required.[21] While possessing various modes on a system is not new, there are new ways to couple increasingly autonomous technologies together in ways that might provide new emergent capabilities. Each sub-component part of the system may not possess select *and* attack capabilities on their own, but when combined, the system of systems does. Thus there may be an implicit loophole for autonomous systems of systems when each individual component does not rise to the level of an “autonomous” system.

[20] James Farrant and Christopher M. Ford, 2017, “Autonomous Weapons and Weapons Reviews: The Second International Weapons Review Forum”*, International Law Studies*, vol. 93, pp. 389–442. Available online at: http://stockton.usnwc.edu/cgi/viewcontent.cgi?article=1710&context=ils.

[21] John K. Hawley, 2017, “Patriot Wars: Automation and the Patriot Air and Missile Defense System”, Center for a New American Security, https://www.cnas.org/publications/reports/patriot-wars.

Optimizing

Many machine learning systems, particularly those which use deep neural networks, work towards optimizing some behaviour. This could be optimizing for time, or monetary value, or even energy usage. For an AWS, humans will need to make conscious choices about what the system is trying to optimize.

First, because such optimization models are mathematical algorithms that rely on vast amounts of data, there will always be two forms of optimization: computational time and computational power. In this way, a system needs to be able to quickly evaluate its present state to determine the correct action. However, if this is a machine learning system, based on something like reinforcement learning, it will determine the correct action by evaluating all possible actions, then choosing the action from this set that will maximize its future reward. Rewards are like utility functions, defined as some set of good or goal. This could be points in a game, or even correctly grasping an object. Whatever the reward, and thus the reward function, this is accumulated over time after many training experiences where the AI explores and tries actions over-and-over to learn a value function.

If the environment is very complex and the system is taking in all of this sensory data, planning a course of action, choosing the best way to get there, and then choosing the best target to fire upon, this could become very computationally complex. This would then entail an increase in computational time.

Second, optimization problems could result in one objective being pursued relentlessly despite other common-sense values being salient. For example, if one wanted to use a "cleanliness" measure as what to optimize in an autonomous vacuum cleaner, it may just keep vacuuming up one space and dumping out the same bit of dirt indefinitely because it will maximize its reward (i.e. pick up dirt) faster this way. Or perhaps it finds a way to break its sensors so that it cannot perceive dirt and thus finds a way to hack into its reward function.

Third, what would we be asking an autonomous weapon system to maximize or optimize? As it is a weapons system, are we looking at how many military objectives it can attain, how many lives lost, how many lives saved, how much it can minimize collateral damage, or maximizing the most damage with the least amount of energy or explosive ordinance? If one wanted to take all of these considerations into the equation then there will be cases of trade-offs, conflicting values, or inadequate resolutions through satisficing. For example, if an AWS has three values it is trying to optimize, but it finds that it cannot satisfy all of them given the situation, we might tell it to find a mathematical mean or average of the three and then take that course of action. However, while mathematically a simple way to resolve a conflict, in practice, such a conclusion may be worse, all things considered.

Scrutability/explainability

The decision process for machine learning systems, particularly those that rely on deep neural networks, is extremely difficult to scrutinize. In essence, when a neural network has many layers, with each layer consisting of various nodes, the connections between the layers and the nodes become so complex that it is almost impossible to understand how the system came up with its "output" or decision. Users may thus not know why a system classified a certain object the way it has, or whether there is an error somewhere in the system producing spurious results. If one were to try to observe the system working, it would be too complicated to trace back through the layers to figure out exactly what went wrong and where. To address this problem, recent attention has focused on generating "explainable AI", in the sense that it would be possible to generate both an explanation of the model and an interface that provides a user with some form of understandable explanation of why the algorithm produced the decision it has.[22]

However, this work so far is still in its infancy. Moreover, the more complex the task, the more interconnections between different neural nets required to complete the task, and so the more computationally complex this becomes. Due to these difficulties, this is why it has become commonplace to call these systems “black boxes”.

Lethality

The discussion thus far in the CCW context has concerned *lethal* autonomous weapons systems. However, this framing omits concerns about the development of autonomous “less than lethal” weapons, as well as concerns about increasingly autonomous anti-materiel weapons.

Lethal anti-personnel weapons systems are often cited as the central concern. These may take the form of either crude identification systems, such as the Super aEgis 2 that can lock onto a human-sized target using infrared sensors at up to a two kilometre distance, or future systems that are more advanced, such as one utilizing facial recognition or behaviour recognition as an indication of combatancy. These types of anti-personnel systems pull us in different directions. On the one hand, too crude systems may fail to meet the necessary principle of distinction. On the other, the deployment of extremely advanced systems may feel more like targeted assassination or individualized warfare. Additionally, concerns might emerge that were a less-than-lethal autonomous weapon with either crude or advanced capabilities to be created, it would be relatively easy to modify the weapon to project lethal force.

Of course, in the same way that existing anti-materiel weapons target objects yet cause collateral deaths, increasingly autonomous anti-material weapons will likely also cause death as a secondary or collateral effect. For example, weapons that are able to find particular buildings, radar signatures or objects will have to rely on a host of other intelligence data to meet obligations of precaution, necessity, proportionality and discrimination. Yet even if these technical considerations are overcome, there is a subsequent concern that it would be relatively easy to modify—either intentionally by the user or by an adversary—these systems to attack classes of targets outside of their normal set.

Predictability and reliability

Often, the most cited characteristic of an “automated” weapon is that it is predictable, whereas an autonomous weapon is said to be “unpredictable”. The notion of autonomy carries with it an ability to change courses of action or make other decisions than were initiated by a human operator. Automation is often likened to some routine or being mechanistic and directly linked causally from an input, such as a command, to an output action. Autonomy, however, carries with it a notion of freedom of action, where the causality of an act is not directly linked from input to output, or it has the potential to have intervening agency from input to output.

Predictability is the state of knowing what will occur in the future, given the current or present state of affairs. Being unpredictable is, then, the inability to know what will occur in the future, even if the state of affairs remains the same. In technical terms, predictability is the ability to know that a system will act with a high degree of probability in a particular way at a particular time.

How one achieves this knowledge can be quite rudimentary, such as through simple mechanisms and physics (e.g. a landmine). Or it could be quite complex, such as a machine learning algorithm

[22] See, for example, https://www.darpa.mil/program/explainable-artificial-intelligence.

that has millions of data inputs and hundreds of layers in a neural net. In either case, the knowledge of the future event would be considered a probability distribution given the inputs from the environment.

Reliability, however, is quite different from predictability. Prediction is the ability to know the future environments, states or actions (or pairs of them). Reliability is more akin to consistency. That something performs reliably means that it acts in ways that are expected, has a history of acting in accordance with prior, or expected, patterns of actions or behaviours and performs consistently well. For instance, a laptop user has an expectation that her computer will turn on when she lifts the lid, because in her past experience it always has done so. In this way reliability and predictability are often linked. She is able to predict based on the consistent performance from past experiences. However, when conditions radically change, such as if she were to leave her computer in the sun and the battery dies quickly, then her computer may not turn on when she opens it. In this sense, if she was unaware of the consequences of leaving the computer in the sun, then one would consider its failure to turn on a sign of unreliability. Or perhaps from past experiences, she knows that leaving her computer in the sun will result in the battery draining. And so, if it does not turn on, she can recall prior experience to explain why its unreliability is occurring. Reliability is the expectation of consistent performance. Thus one could have very predictable but highly unreliable systems.

Precision and accuracy

Precision is the degree to which things cluster together. These can be measurements taken in the same place or salvos of weapons all landing on the same spot or near to it. For example, if one were to fire arrows at a target, they may all hit the same spot in the outer ring. This would be precise, however, it would not be accurate because the intended or “correct” target would be the “bullseye”.

Often we talk of “precision guided munitions” as those systems that are highly precise. Technically, these systems are capable of Going onto an Object in Space (GOIS) or Going onto a Target (GOT). In essence, the “smart bomb” is guided by something to locate a designated target. In the case of GOIS, this is typically a geo-positioned satellite that guides the system to a predesignated target coordinate. Whatever is at that coordinate in space and time becomes the object of attack. With the GOT system, the munition is able to get to a particular space, but once there, is looking for a particular type or class of target (say a ship, tank, or radar emitter). In sophisticated systems, we may have both capabilities in a weapons system. Moreover, in some new missile systems, such as the Long Range Anti-Ship Missile (L-RASM), they are able to coordinate amongst a fired salvo to triage which munition will attack which vessel or where.

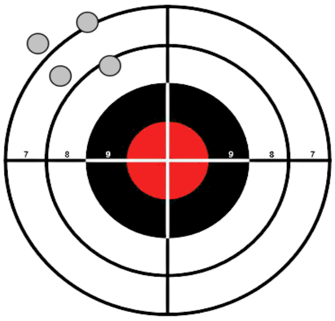
AWS may or may not be very precise because the technology that enhances precision can vary on munitions or platforms, such as the type of guidance system present, whether there are serious measurement errors, or composition of the physical environment. Or one could merely have autonomous “dumb bombs”.

What is crucial for the discussion is to note that it is not about precision, it is about *decision*. In the case of an AWS, there is both the ability to get to a particular time or location in space, but the *choice of target is up to the weapon system.* In essence, precision guided munitions are very exact in carrying out a previous decision made by a human operator or commander; if made more autonomous, precision guided munitions would also have the ability to not just find their way to the target but decide which targets and whether to engage. In our archery case above, the AWS decides whether the bullseye is really the target in the first place; where it puts its arrows is a different matter. *Accuracy* is different than precision. It is the “location of the point of impact for a given aim point on the target."[23] In essence, accuracy is measured by the distance between the aim-point of the selected and intended target and where the munition actually strikes. Thus a precise weapon can be used inaccurately.

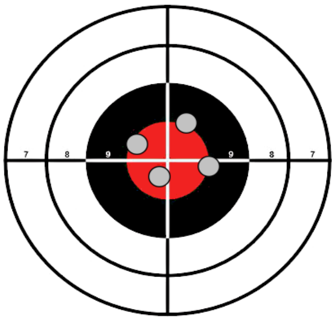
Additionally, accuracy carries with it the idea of correctness or completeness. For instance, according to the US Department of Defense, during the process of Combat Identification (CID), one is to have "accurately characterized all the detected objects in the operational environment sufficient to support an engagement decision." In this sense, it would be the correct correlation between identification of civilians and civilian objects and identification between friend and foe.

Accuracy, then, has the potential to increase or decrease due to noise, data, sensor links, etc. In autonomy at rest systems, say battle management software or decision aides, autonomous systems may be able to integrate large amounts of sensor data, natural language voice processing, images, and prior courses of action. If the system performs as intended, then the autonomous system may increase situational awareness and provide greater accuracy. If, however, there are unforeseen failures through, for example, the integration of multiple sources of data or intelligence, then this may decrease the accuracy. Accuracy, then, is also about the right, or correct, decision or selection.

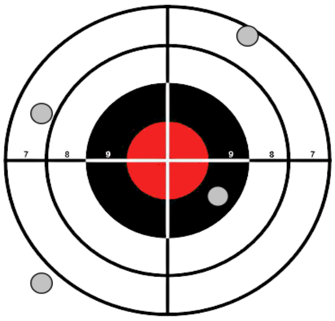
To illustrate:



Precise, but not accurate



Accurate and precise



Neither accurate nor precise

Accountability

In a traditional military organization, accountability is hierarchical. Soldiers are trained to obey orders and strict procedures, and are accountable to their commanders, who are responsible for making decisions and providing leadership. Accountability flows upward, through the chain of command. The higher up on the chain, the more power one has to make various decisions and order events. For military operations, accountability for AWS remains an unclear area.

One may claim that delegating an order to the machine does not absolve the individual who made the order. However, if one delegates an order to an AWS where the AWS must make many "decisions" about how to carry out that order without input or intervention from a human operator, then there is less direct causality between the time of order and the completion of the task. This intervening agency from an AWS may actually flatten out some of the hierarchical structures in military organizations. If this is so, then accountability may also flatten here. Thus the organization,

[23] J.S. Przemieniecki, 2000, *Mathematical Methods in Defense Analyses*, vol. 11, (3rd edition): "In any given weapon system, however, the most important characteristic parameters are its accuracy, effectiveness against a specific target, and reliability. These parameters are somewhat related, because, as the weapon's accuracy decreases, so does its effectiveness."

which relies on hierarchy of command and responsibility, moves more decision-making capabilities to a non-human agent that cannot be held accountable.

One may also claim that it does not matter who "pushed the button" or gave the order, accountability always flows upward to the leaders and ultimately the State itself. *De jure*, there is someone or something accountable. However, if the actions of an AWS are unforeseen, unintended and deemed an "accident" this may engender an odd situation where operators, commanders and the State are all absolved of accountability because each malfunction is always an accident. In short, the introduction of AWS into the battlespace changes that space to one where intentionality and accountability are *de facto* not possible.

Human decisions and agnostic machines

Presently autonomous systems do not understand concepts, and they do not have "feelings". Irrespective of what sort of AI architecture runs an autonomous system, that system does not have higher-level preferences about where it is used, when or by whom. It will attempt to carry out its order and fulfil its tasks regardless of the environment it finds itself in. In this way, it is "agnostic", and can be—although might ought not to be—used in any situation. We may be able to design systems that are able to produce confidence measures about what objects it sees, or whether or not it is in the correct environment for its use, but even these confidence measures will merely be mechanisms for humans to place thresholds on various courses of action. The system itself is agnostic about whether it requires a 10 or .001 uncertainty measure since it is the human, not the system, that determines what levels of uncertainty are acceptable. Moreover, we cannot rely on the system to "correct" us because humans are the ones telling it how much margin of error or uncertainty we are willing to accept.

Even if systems are correctly able to identify increasingly grey or fuzzy targets with great clarity, expanding a system's area of application may in fact be highly unstable. Because systems will attempt to carry out orders regardless of environmental changes, there may be an increasing distributional shift from what a system was trained on to what it encounters in live environments. As the system does not understand, it merely attempts to do what it was initially trained to do. Even if we attempt to build in a failsafe, this may not in fact work if the machine cannot determine its own uncertainty.

Systems of systems

"Systems of Systems" (SoS) describes the composition of component parts that are each individually considered a system.[24] A "subsystem must be able to function independently on occasion, and yet be a cog in a larger machine on other occasions. Dynamics in the evolving structure is a peculiarity of SoS."[25] Noteworthy of SoS is that they exhibit emergent behaviour, which is behaviour that is not predictable in advance. For example, individual sub-component parts of the system may not possess select *and* attack capabilities on their own, but when combined, the system of systems does.

> For agents that are capable of autonomous or semiautonomous operation, cooperation and collaboration imply task level interactions. Indeed, in the SoS context, it should be expected that component systems have their imposed goals but might generate (in an

[24] Tariq Samad and Thomas Parisini, 2011, "Systems of Systems" in T. Samad and A.M. Annaswamy (eds.), *The Impact of Control Technology*, http://ieeecss.org/sites/ieeecss.org/files/documents/IoCT-Part3-04SystemsOfSystems.pdf.
[25] Ibid.

> evolutionary way) their own goals—causing dynamic interactions with other component systems.[26]

Thus there may be an implicit loophole for SoS when each individual component does not rise to the level of an "autonomous" system, but the emergent and dynamic behaviour of the SoS is itself autonomous.

Communications and connectivity

Connectivity is a characteristic frequently raised in discussions about autonomy. However, it is not a characteristic of autonomy as such. Rather, it is a justification for developing and deploying autonomous systems in areas where there is likely denied communications, such as underwater, in space, or in communications-denied environments. In a similar vein, some States consider that increasing autonomy is needed in order to operate in so-called "degraded" environments. Increasing autonomy might permit operations in these environments where, until now, operations were limited due to the need for connectivity. Moreover, increasingly autonomous technologies that can operate without active communication might be militarily useful for particular activities, such as stealth operations, or operating in environments in which communications are expected to be jammed. However, there is no technological impediment to autonomous systems having connectivity, or prescribing how, under which conditions or how frequently, the system connects. Rather it is a conscious design decision whether or not to include it as a feature in a system.

Mobility

Often the notion of mobility arises in the discussion of autonomous systems. Mobility is only a characteristic of autonomy in motion systems and not systems that possess autonomy at rest.

Mobility is the ability to move about in an area, which is not inherently problematic. However, in the context of hostilities, the concern is that there can be very rapid environmental changes, and due to this rapid change, there may be drastic differences from a point of deployment to the point of impact. Such a change in environment might lead to such consequences for two reasons. First, it may be that a system that is "frozen" and not continually learning may encounter new data input that it has never seen before. This may mean that it lacks an adequate representation of the world or that its model/training no longer fits. Second, it may mean that a system that continues to learn, thereby avoiding the "frozen" problem, can still suffer from "distributional shift". That is, whether a machine learning system can indicate when it knows that it *does not know* something. Or would it make a confident determination despite the fact that it has never encountered that type of information or environment before?

Speed

One often-cited driver for increasingly autonomous systems is the need for systems that can act "beyond human reaction times"—systems that can act or react in picoseconds. This justification is most easily understood when considering defensive systems, such as a missile defence system, counter-rocket and mortar systems, active defensive laser weapons, defensive space-based systems, and cyber-security "active defence". While it is certainly true that speed is what will give any one party dominance over another, in many of these domains, speed *qua* speed is not a characteristic of autonomy. Rather, speed becomes a strategic advantage, and thus operates as a

[26] Ibid.

justification for why greater autonomy ought to be employed. In this way, speed also becomes a driver for increasing autonomy.

Concerns have been raised that in such rapid defences one may accidentally escalate a conflict, as the attacking side may also have AWS. In this case, there is need for greater sentience, or "understanding", about the situation and broader strategic contexts because more cognitively competent systems may provide an "off ramp" or be able to de-escalate a situation by not immediately engaging. Thus, there are greater incentives to ensure higher orders of cognitive capabilities, either by teaming the system with a human to provide greater situational awareness, or by relying on more complex AI.

3. Definitional approaches

When establishing the mandate for the GGE, the High Contracting Parties to the CCW allocated themselves the task of identifying characteristics of LAWS and elaborating a working definition.

Numerous definitions of autonomous weapon systems—proposed by States, international organizations, civil society groups, and academics—are already in circulation. No two definitions are exactly alike. Despite there not being a common definition of what is an AWS, as of October 2017, 19 States have declared their support for a pre-emptive ban, and many more have endorsed the concept of "meaningful human control" over weapon systems, again with no shared definition.

In the CCW, we have seen the emergence of three main approaches to the question of a definition.[27]

Technology-centric approach

Some governments have expressed a desire to search for a **technical definition** in which **a physical object is described.** This is in line with the way that CCW has traditionally approached specific classes of weapons, focusing on aspects such as technical specifications, range, payload, and intended operating environment. Frequently the CCW has focused on "splitting the category" of systems under consideration into "good" and "bad". Any negotiation thus becomes, in effect, bargaining between what is "in" and what is "out" based on technical criteria devised by users and possessors—sometimes at the expense of addressing the problematic effects of the weapon's use.[28]

Due to this experience, it is perhaps not surprising that some governments approach the LAWS discussion with the desire to draw clear distinctions between autonomous objects and automatic ones based solely on technical characteristics.

However, the tech-centric approach is not without challenges, especially since so much of the discussion of future LAWS systems is speculative.

First is the most practical concern: increasing levels of autonomy could be applied to every weapon system, just as increasing levels of automation have been applied in previous generations of weapons. Autonomy is a characteristic, not a thing in and of itself. It could be applied to any weapon system. It could be applied to different parts of any system—for example, something might be able to determine its path and navigate autonomously, but once on target, humans are involved in the decision to engage. You might have an adjustable object with an autonomous mode, automatic mode and human-operated mode. It will be difficult to capture the variety of characteristics, in various combinations, in a tech-centric definition.

Additionally, the pace of technological development and innovation will quickly date a definition that describes what is "state of the art" today. Such a definition would require constant monitoring and review—and new applications of concern could "slip past" before a monitoring and review mechanism makes a determination. It would be a perverse consequence of the financial challenges facing some of the treaty regimes today were financial hardships to impede regular review and updating of the treaty—and thus perhaps letting "slip past" worrisome applications or developments.

[27] For a similar analysis of categorization of definitional approaches, see Vincent Boulanin, 2016, *Mapping The Development of Autonomy in Weapon Systems: A Primer on Autonomy,* SIPRI, pp. 29–30, https://www.sipri.org/sites/default/files/Mapping-development-autonomy-in-weapon-systems.pdf.

[28] See, for example, John Borrie, *Unacceptable Harm: The History of How the Treaty to Ban Cluster Munitions was Won,* UNIDIR, p. 333.

Secondly, ultimately a definition will draw the line between what is to be regulated and what is not—yet in cases where there is a significant area of differing understanding, there is considerable space for manoeuvre to adhere to the letter without adhering to the spirit of the definition. For example, governments might decide that autonomy in weapon systems in defensive applications such as Iron Dome, C-RAM, THAAD, or perhaps even armed sentry patrol bots in very limited geographical areas that are demarcated as "no man's land" are acceptable. Rather it is "offensive" systems that require a stricter regulatory approach. But an offensive and defensive system are physically identical—one simply modifies the conditions under which it is permitted to engage. For instance, with laser weapon systems, one might as easily use the directed energy weapon to engage an incoming threat, such as an unmanned aerial vehicle, or it may be used to target other objects outside of its authorized use as a purely defensive system. Or an autonomous platform for non-weaponized applications—such as surveillance or reconnaissance—could be rapidly weaponized either deliberately or in an ad hoc manner. UAVs started out as a surveillance platform, only to be weaponized later.

And a tech-centric definition would need also to take into account that militaries are interested in a wide variety of increasingly autonomous objects, such as autonomous supply convoys, not just weapon systems. What sort of protective measures would such objects be equipped with? For example, in the case of the autonomous supply convoy, one certainly would want to be able to ward off an adversary's attempts to capture supplies—these could be food, equipment, or even weapon stocks—so it would be logical to equip these autonomous convoys with a self-defence system. So now one has a mobile autonomous object with some sort of a defensive weapon mounted on it, that might not be captured in the definition of an AWS because it is not primarily *designed* to be a weapon system, let alone an offensive system—yet it essentially can do the same thing. It all has all the same hardware, the sensors, and an effector in the form of a weapon.

Lastly, as noted in UNIDIR's first Observation Reports,[29] a tech-centric frame also tends to be exclusionary, as many governments don't feel technologically "fluent" enough to participate in the conversation.

Human-centred approach

A second approach is to describe an **AWS in relation to a human user**. Increased autonomy means by definition that a human has delegated some level of control/decision-making to an object. In CCW this approach of describing the role of the user is where we have heard talk of humans being in, on, or out of the loop, as well as emergence of the concept of "meaningful human control" or in the terms of the US Department of Defense directive "appropriate levels of human judgement". This approach is grounded in existing legal commitments and norms, and it is easier to participate in regardless of one's level of technological sophistication.

A human-centric definitional approach:[30]

- provides a common language for discussion that is accessible to a broad range of governments and publics regardless of their degree of technical knowledge;
- focuses on the shared objective of maintaining some form of control over all weapon systems;

[29] See UNIDIR, 2014, *Framing Discussions on the Weaponization of Increasingly Autonomous Technologies,* UNIDIR Resources no. 1, pp. 7–8; and UNIDIR, 2014, *The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward*, UNIDIR Resources no. 2, p. 3.

[30] Drawn from UNIDIR, 2014, *The Weaponization of Increasingly Autonomous Technologies: Considering how Meaningful Human Control might move the discussion forward*, UNIDIR Resources no. 2, p. 3.

- is consistent with IHL regulating the use of weapons in armed conflict, which implicitly entails a certain level of human judgment and explicitly assigns responsibility for decisions made; and
- is a concept broad enough to integrate consideration of ethics, human-machine interaction and the "dictates of the public conscience", which can be marginalized in approaches that narrowly consider just technology or just law.

However, a human-centric approach may be impractical from a technological perspective, as there is much evidence about how automation changes the relationship between humans and these systems in negative or costly ways.[31] In addition, it may be difficult to test, evaluate or verify if a human is considered part of the system.

Task/Functions approach

A third definitional approach focuses on identifying the tasks or functions delegated to a weapon that makes it autonomous. For example, the International Committee of the Red Cross (ICRC) has taken the approach that an autonomous weapon is one that possesses autonomy in its "critical functions" where these functions are specific to selecting (i.e. search for or detect, identify, track, select) and attacking (i.e. use force against, neutralize, damage or destroy) targets without human intervention.[32]

The functionalist approach is without prejudice as to whether such systems ought to be regulated as such. It is a broad approach, being inclusive of both previously considered "automatic" weapons in use and potential future systems.

The functionalist approach:

- is not reliant on a particular kind of technology or state of technological development;
- is for the most part focused on the particular activities of selecting and attacking; and
- is sufficiently simple to reach broad agreement.

Some may claim that a functional approach may be too broad, "capturing" existing systems in the definition. Further, it may be overly inclusive due to the nature of contemporary armed conflict because multiple subsystems that are not attached to a weapon's platform are utilized for "selecting" targets.

Sequencing the approaches

Until now, these three definitional approaches have been seemingly competing for primacy. However, in fact, they are complementary if sequenced correctly. Starting with a human-centric approach allows us to reaffirm human responsibilities and existing legal frameworks regardless of the specific new technology. Then, turning to identifying the key/critical features or tasks that we have uncertainties or concerns about when autonomy is applied to them will help narrow down the scope of the discussion. Finally, after determining the appropriate and necessary human role, as well as the tasks of concern, one can turn to a tech-centric conversation.

[31] See, for example, John K. Hawley, 2017, "Patriot Wars: Automation and the Patriot Air and Missile Defense System" Center for a New American Security, https://www.cnas.org/publications/reports/patriot-wars.

[32] See, for example, ICRC statement of 13 May 2014 at www.icrc.org/eng/resources/documents/statement/2014/05- 13-autonomous-weapons-statement.htm; see also ICRC, 2014, *Report of the ICRC Expert Meeting on Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*, www.icrc.org/ eng/assets/files/2014/expert-meeting-autonomous-weapons-icrc-report-2014-05-09.pdf.

Were governments to first explicitly agree on what role they want to maintain for humans (for operational, legal and ethical reasons), and then to identify the tasks where autonomy might call these roles into questions, States would then be able to describe what technological features/characteristics they would need to see/avoid in future weapon systems. Then they would be able to determine appropriate regulatory responses.

It is natural that proponents and opponents of AWS will seek to establish a definition that serves their aims and interests. The definitional discussion will not be a value-neutral discussion of facts, but ultimately one driven by political and strategic motivations. To mitigate this as much as possible in the early stages of discussion, it is essential to separate the definition that is a description of a category from the possible subset of that category that requires a regulatory response. States must be clear about whether the definitional exercise in CCW is to define a larger category of AWS (of which a subset might be problematic or raise particular concerns) or just to define the subset of potentially problematic applications. For some, these two categories may have very little overlap, while others see them as a near eclipse.

Some States seem reluctant to engage in the broader definitional exercise, perhaps fearing that agreeing to a wide and encompassing definition would capture existing or near-term systems and thereby call into question their legitimacy or legality. *The definition discussion is different than, and should not be confused with, the categories that States might decide to eventually regulate or control.* A logical approach to advance the definition discussion would be to:

- First, capture all possible autonomous systems in a broader definition;
- Second, within that broad definition, identify the potentially problematic applications; and
- Third, determine what are the appropriate policy responses to the second category.

4. Sample working definitions

What follows is a selection of working definitions that have been put forth by governments and other stakeholders in the autonomous weapon discussion. This section is in no way exhaustive. These definitions were chosen as illustrations, and even these are likely to evolve as a result of domestic and international discussions. Rather, it is a selection used to illustrate how particular approaches attend to certain concerns or characteristics and not others. By presenting these here we offer governments an opportunity to consider how to further refine or improve upon proposed definitions as the GGE moves forward.

Government of the Netherlands

Autonomous Weapon System: "A weapon that, without human intervention, selects and engages targets matching certain predefined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention."[33]

The Dutch working definition narrows the scope of what constitutes an AWS by requiring not only weapon systems that can select and attack without requiring human guidance or intervention, but also that these systems cannot be recalled or stopped after deployment or launch. The justification for this narrowing scope is that the Dutch government believes that "meaningful human control in the wider loop" still governs the "wider targeting process". As long as humans are preselecting the criteria on which weapons make targeting decisions at the time of attack, as well as that humans make considerations about "aspects such as target selection, weapon selection and implementation planning (time and space), an assessment of potential collateral damage" and "battle damage assessment", the system would be considered permissible and presents no "additional ethical issues compared to other weapon systems".[34]

The Dutch working definition stresses the need for human engagement and accountability. The focus on multiple time frames, such as weapon design and testing to engagement and post-attack assessment, is correct. Indeed, because the Dutch note the obligations for humans at each time phase, it appears to fall within the human-centric approach noted above. It also reaffirms existing IHL obligations on both individual commanders and States, such as for States to comply with Article 36 weapons reviews.

While "meaningful human control in the wider loop" still governs the "wider targeting process", the working definition does not mention meaningful human control. As such, marrying Dutch support for meaningful human control to the working definition may be difficult once other States enter the discussion on the definition.

The Dutch definition is very narrow, limiting the discussion to systems that select and engage targets without human intervention *and* cannot be stopped by humans. It seems to imply, then, that weapon systems that select and attack without human intervention, but could be recalled or stopped, would not be autonomous weapons. This may restrict the label of "autonomous weapon system" to very few systems, such as swarms or autonomous submersibles without communications. Yet, as the Dutch government notes in its working paper, "even if it became technologically feasible, there seems to be no reason why a State would have the ambition to

[33] Government of the Netherlands, 2017, "Examination of Various Dimensions of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, in the Context of the Objectives and Purposes of the Convention", document CCW/GGE.1/2017/WP.2 of 9 October.

[34] Ibid.

develop a weapon system that is intrinsically not under human control."[35] Though, if States are developing and launching systems that cannot be stopped, it would seem that at minimum a large degree of control is lost.

The concept of "wider loop" could benefit from further conceptual clarity, as the paper presumes a "narrow loop", yet does not describe which tasks are delegated to the system in the "narrow loop". The Dutch paper notes a "prominent role for humans" in programming target characteristics, target and weapon selection, elements of planning and assessment of potential collateral damage, as well as Battle Damage Assessment.

There will be antecedent design decisions made by humans, and there will be a decision by someone to deploy an AWS. There is no technological impediment that ensures that other decisions noted as part the 'wider loop' will continue to be made by humans in the future. For example, if a system can choose—that is select—and attack a target without human intervention, the system will require various navigation, planning, sensing and engagement-related capabilities. As an example, the current F-35 already has limited local battle damage assessment capabilities for it to be able to function with its pilot.

Finally, it is unclear how the Netherlands would like to address governance of AWS. On the one hand, it states that there are no new ethical concerns for fully autonomous systems under meaningful human control in the wider loop. On the other hand, the paper does not explicitly reference that fully autonomous weapon systems *without* meaningful human control require regulation. Rather, they state that they do "not support a moratorium on the development of fully autonomous weapon systems", citing the difficulty of regulating the dual-use nature of AI.

Government of France

> "Lethal autonomous weapons are fully autonomous systems. [...] LAWS should be understood as implying a total absence of human supervision, meaning there is absolutely no link (communication or control) with the military chain of command. [...] The delivery platform of a LAWS would be capable of moving, adapting to its land, marine or aerial environments and targeting and firing a lethal effector (bullet, missile, bomb, etc.) without any kind of human intervention or validation. [...] LAWS would most likely possess self-learning capabilities."[1]

The French working definition in part circumscribes areas of what LAWS *are not*. They are not:

- existing automatic systems;
- linked in any form of communication or control to "the military chain of command";
- supervised in any way, or capable of "human intervention or validation";
- liaised with "the weapons system";
- able to provide "permanent and accurate situational awareness and the operational control" to the commander; or
- predictable.

The benefits of the French approach are in the specificity of which types of systems ought to be considered as "autonomous", while also indirectly providing a fuller account of what it takes "autonomy" to mean. By precluding automatic systems from discussion, and by definition any systems that are non-lethal or less-than-lethal, the definition suggests a bright line distinction for

[35] Ibid.

autonomy. Systems that are pre-programmed to act in a particular manner without any freedom of adaptation, variation or discretion would be considered as automatic, not autonomous.

Furthermore, the definition and accompanying discussion also hints at what may be required for the "selection" of targets. "LAWS" it suggests "would most likely possess self-learning capabilities" because the complexity and diversity of potential military scenarios could not be "pre-programmed". The system would need to "learn," and the "delivery platform", which could ostensibly be separate from other weapon system components, "would be capable of moving, adapting, […] and targeting and firing a lethal effector." This learning, France suggests, would mean that "the delivery system would be capable of selecting a target independently from the criteria that have been predefined during the programming phase, in full compliance with IHL requirements." This wording, however, implies that only systems that continue to learn once deployed would be considered autonomous, and that any other systems possessing machine learning but not continuing to learn once deployed would be considered as automatic.

The 2016 French definition focuses on the far end of the autonomy-capabilities spectrum, excluding, for example supervised "autonomous" systems. In the French paper, any supervision—even of a system that can act independently and without human intervention—is excluded from the concept and definition of autonomy. Supervision necessitates some form of communications link (whether unidirectional from user to the object, or bi-directional with the object being able to communicate to the user as well). While it is certainly feasible that some forms of autonomous systems will operate without a communication link (at least in some circumstances), some might claim that this may prove an overly restrictive requirement for a system to be considered "fully" autonomous. Additionally, it appears that systems that operate for extended periods without communication but may "check in" with commanders would be excluded from being autonomous, as would any systems that have multiple modes, or continuums, of autonomous behaviour.

The definition may preclude consideration of systems that may be comprised of many subcomponent parts or munitions, each of which is not deemed "autonomous" in isolation, but by their use together the emergent behaviour appears to be autonomous. In these cases, there may be a "total lack of human supervision" at the time of attack, but not during any planning or initial deployment stages. For example, a swarm of micro-drones may fall under the heading of "automatic" in this definition, but acting in concert they may exhibit emergent behaviour. It is unclear whether those types of systems or modular weapons systems would qualify as AWS under the French definition.

France's definition uses the phrase "lethal autonomous weapon system", implying that the weapon system must be directed towards human targets, as it is a *lethal* weapons system, and therefore not applying to anti-materiel weapons, countermeasure systems, or non-kinetic systems. It does not address whether permissibility of such lethal systems may rest on whether they are for purely defensive purposes, such as perimeter defence. France's definition raises some crucial issues about machine learning and design choices. It is true that learning systems are unpredictable, in the sense that they may learn something unforeseen. However, it is a design choice as to whether learning is frozen prior to deployment. Furthermore, the notion that a self-learning system will select targets "independently from the criteria that have been predefined during the programming phase" is not inevitable. Learning systems are trained on a set of data, and how that learning takes place and the technical specifics that go along with it, may entail that the system cannot "select" new targets outside of the training data. It may attempt to fit new knowledge into its model of the world, but that would mean that it is incorrectly identifying some object. This may be due to some unknown relations in the training data and the system was not validated on a set of data previously unseen, or it may be due to uncertainty. What is unpredictable is how the system learns in a given model, and how it will extrapolate that learning to new environments.

Finally, the French approach includes two restrictive requirements: First is how "full autonomy–and the absence of liaison with the weapons system—contradicts the need for permanent and accurate situation awareness and operational control." Further clarity would be beneficial as it is unclear how a weapons *system* could not liaise with itself. If a weapon system is a combination of one or more weapons with all related equipment, material, services, personnel, and means of delivery and deployment required for self-sufficiency, then it will by definition "talk to" or link with itself. Second is the emphasis on the "total absence of human supervision". States and militaries already have the option to deploy weapons systems without human supervision. Fire-and-forget munitions, for example, require no further guidance after launch, and in many instances, do not need to be observed by a human operator. Likewise, some cruise missiles already possess automatic target recognition software and do not require guidance, control, or supervision by humans. While most militaries will observe the weapons during flight and upon detonation, this is to keep commanders aware of battlespace changes and is not a requirement under IHL.

International Committee of the Red Cross

Autonomous Weapon System: "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."[1]

The ICRC's working definition for an AWS takes a functionalist approach. The definition considers the technical, legal and ethical requirements for "control" and, subsequently, human-machine interaction. This functionalist approach does not prejudice which functions are or are not problematic. Rather, it states that any system that can select (with whichever capabilities the system requires for selection) *and* attack (with whichever means, methods or munitions the system deploys) without intervention by a human operator qualifies as an autonomous weapon system. In that way, the ICRC definition is quite neutral. It attends to the wider category of AWS, not all of which would be necessarily problematic or of concern.

Moreover, since the definition is without prejudice, it does not claim that a system with autonomy is prohibited *per se*. Therefore, a system may permissibly have autonomy in its critical functions, so long as it complies with international humanitarian law obligations (such as discrimination, proportionality, and precaution).

However, to ensure that all new means and methods of war are compliant, States are required to undertake legal reviews under Article 36 of Additional Protocol I to the Geneva Conventions. Given this obligation, the ICRC's definition provides further, though secondary, support for Article 36 obligations and also subsequent obligations for additional life-long testing and certifications for any AWS. As the ICRC notes:

> The ability to carry out [an Article 36] 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.[36]

As the ICRC states in its 2016 working paper, "a certain level of human control over attacks is inherent in, and required to ensure compliance with, the IHL rules of distinction, proportionality and precautions in attack." By noting that States have obligations to comply with IHL, and that militaries

[36] Ibid, p. 81.

cannot field weapons out of control, the ICRC's approach urges States to consider human–machine interaction and permissibility of delegating particular tasks or combinations of tasks within the targeting cycle. "From the ICRC's perspective, a focus on the role of the human in the targeting process and the human machine interface could provide a fruitful avenue for increasing understanding of concerns that may be raised by autonomous weapon systems, rather than a purely technical focus on the 'level of autonomy'."

The ICRC definition does not address many of the questions related to the difference between "automatic" systems versus "autonomous" ones, nor does the definition discuss weapons systems that may have various "modes" that can increase autonomous functionality (adjustment). Since its definition is without prejudice and inclusive, it may encompass systems with varying modes, or even ones that may have emergent capabilities.

The ICRC's language does not provide a definition of autonomy, and so it may want to include automatic and autonomous systems together in the definition, or merely exclude the word autonomous altogether, particularly since the definition is without prejudice to the regulation or prohibition of the *class* of weapons systems. However, if one were to include both automatic and AWS in a definition, without prejudice, then one would have to have some form of agreement on all the critical functions of a weapons system. Implicitly, then, States would recognize that such critical functions would be the same in both automatic and autonomous systems, but that due to technical difficulties in defining exact limits or thresholds, the functions could potentially change in kind or in degree. This would remove the need to define levels of autonomy altogether.

Finally, the wording of "select" in "select and attack" may suffer from circularity. The ICRC includes "detect" and "select," as well as other capabilities (e.g. identify, track) to explain what it means by "select." However, each of these terms are conceptually different, though they may require the same or similar hardware and software technologies. Detecting a target is to sense its presence, but to select it is to choose among potential target objects. Depending upon how one defines "select", one could mean that a human "selects" all the target signatures for a target library and the machine merely matches signatures to the library. Or one could mean that selection occurs at the time of attack and the system is choosing among an array of pre-selected targets. If we define select in the first sense, then rarely—if ever—will any system truly select a target. If we define it in the second sense, then the scope is much broader.

Government of Switzerland

Autonomous Weapon Systems: "Weapons systems that are capable of carrying out tasks governed by IHL in partial or full replacement of a human in the use of force, notably in the targeting cycle."[1]

The Swiss government's working paper suggests a "compliance-based approach" to AWS. This definition seeks to push forward the thinking on autonomous weapons by being as inclusive as possible in the boundaries of what may be considered an AWS. Additionally, the definition expands the scope of potential systems for consideration by not only remaining silent on whether the system is lethal, non-lethal or less-than-lethal, but also whether and to what extent any particular task is carried out by a system. The definition and the working paper that supports it also does not "prejudge the appropriate regulatory response" for AWS.

The strength of Switzerland's approach lies in its inclusivity and its flexibility, as well as how it couples the notion of autonomy to the accomplishment of particular tasks. In terms of inclusivity, the Swiss proposal is explicitly sensitive to "facilitating compliance" and so it encourages the identification of “best practices, technical standards and policy measures” that help to “complement, promote and reinforce” international obligations.

Moreover, the flexibility of the Swiss concept is that it can account for some of the most pressing questions related to AWS, such as what ought to be included in the definition, as well as whether autonomy exists as a dichotomy (automatic or autonomous) or as a continuum. Because the definition looks to “the partial or full replacement of a human in the use of force” it requires States to look at the targeting cycle as a compilation of related tasks. By requiring States to make explicit the assemblage of component parts or tasks in the targeting cycle, whether by human or machine, it may open the door for a variety of kinds and combinations of systems for review. Task-based analysis could then incorporate a variety of subcomponent parts, teams, or integrations.

This tasked-based analysis, moreover, could provide answers to questions pertaining to whether the system is offensive or defensive, anti-materiel or anti-personnel, as well as which functions are “critical” to the task at hand. Functions need not necessarily relate to engagement-related functions, but could also relate to decision aids embedded in weapon systems. For example, as the Swiss paper noted, if an “AWS is expected to perform [a] proportionality assessment by itself, that aspect will need to be added to legal reviews of these systems” (§ 23). If we define a weapon *system* as a combination of one or more weapons with all related equipment, materials, services, and personnel, then, the portion of the system that is completing the task of proportionality assessment, as a decision aid to the operator, who then chooses to fire or launch a weapon, would be under the need for assessment because that task has been delegated to the component part and not the human. Compliance with proportionality-related tasks, then, requires analysis of how proportionality subcomponent part functions, as well as how the output from that component may influence or affect other subcomponent parts (such as through human factors analysis).

One may consider the compliance-based-approach to be too inclusive, as it seems to admit that any weapon system that utilizes information communication technologies constitute AWS. As the definition states that AWS simply are “weapons systems that are capable of carrying out tasks governed by IHL in partial or full replacement of a human in the use of force, notably in the targeting cycle” (§ 6). Since most weapon systems today utilize some form of an information communication technology to complete some portion of a task previously performed by a human during the use of force, it would seem to imply that almost all present-day systems are AWS, or else that further precision is necessary to narrow down how ICT use in an autonomous system differs from its use in existing systems.

For example, even if a military did not utilize a precision guided munition that would take over tasks related to detecting a target object and employing force against it, and instead utilized a “dumb” munition during the target engagement phase of an attack, the problem is that if any subcomponent parts were automated in the weaponeering or capabilities analyses, then by the definition of a weapon system above, it would appear that due to “the partial replacement of a human in the use of force, notably in the targeting cycle”, this system could be considered as an AWS. This conclusion, however, may go against assertions that there are no presently existing AWS.

Lastly, the Swiss working paper appears to support the notion that non-kinetic effects, such as cyber operations, will be eventually recognized as a use of force under IHL.

Government of the United Kingdom

> “An autonomous system is capable of understanding higher-level intent and direction. From this understanding and its perception of its environment, such a system is able to take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous unmanned aircraft will be predictable, individual actions may not be.”[1]

The UK provides its most recent account of autonomous weapons in conceptually robust terms. Like the ICRC’s position, the UK focuses more on the required cognitive capabilities of an AWS rather than on critical functions. The UK definition emphasizes:

- understanding human intent;
- context awareness and sensory perception; and
- goal-orientation/purposiveness.

The strengths of the UK approach lie in its forward thinking about the potential abilities of AI at work in the future battlespace. In this way, the UK is looking toward how humans and AI can interact and share “mental models” of the world and each other. That is, they would both be able to understand the other’s actions, goals, intent and reasoning.

Additionally, the UK’s definition identifies that autonomy is also about decision-making[37] capabilities, and that an autonomous system can take decisions on its own from a variety of courses of action, without human oversight or control. These could be constrained to particular contexts or tasks, but emphasizing the decision-making capacity is central to a definition of autonomy. This capacity, moreover, is not affected by the presence or absence of a human observing the actions of the system.

Due to the UK’s insistence on robust cognitive capabilities, the definition seeks to demarcate a bright line between automatic systems, which are described as pre-programmed and predictable. This seems to also imply a difference in the ability to choose or “decide” targets, where automatic systems may instead detect previously chosen or designated ones.

In the UK approach it is unclear how to test that a weapons system “understands” and possesses an “appreciation of commander’s intent,” or can understand “why” a human ordered it to do a particular task or action. These more complex cognitive abilities would require an autonomous weapon to possess the ability to understand concepts. However, it is unclear whether this is technologically feasible given the current state of the art in AI. Despite significant advances in natural language processing, getting AIs to learn the meaning of language, as well as nonverbal cues, and social convention, robust “intent recognition” are yet unresolved in AI research and is likely to remain so for some time.

In a related vein, because the threshold for autonomy in this definition is quite high, it is unclear what is or is not included in the implied definition of “automatic”. Or whether systems that possess various modes for autonomous action are to be deemed automatic or autonomous, or how they should be evaluated (such as according to their highest level of capability or on each discrete level). As the Joint Doctrine Publication utilizes the phrase “is programmed to logically follow a predefined set of rules with predictable outcomes” to describe an automated system, it is unclear whether the

[37] “[A]n automated weapon system is capable of carrying out complicated tasks but is incapable of complex decision making.” Ibid., p. 42.

UK definition would address hybrid systems that mix more rule-based algorithms with learning systems, or whether "defined rules" would include systems such as supervised learning systems. The definition may also imply single or unitary intelligence, and not the functional equivalent of unintended, unforeseen or emergent autonomy arising from the integration or mixing of various modular "automatic" systems into a SoS.

Questions remain about the extent of learning or goal formation and whether these would be considered automatic or autonomous systems. If top-level goals are given by a commander, but sub-level goals may be formulated by the system, is this sufficient to make the system "autonomous" or is it to be regarded as "automatic"? For example, if the engagement-related function is pre-programmed to a particular space in location and time, but the target-objects within that space are not preselected, is this system deemed autonomous in its target engagement functions but not in its higher-level cognitive capabilities for understanding context and commander's intent?

The definition does not address the word lethal, or whether discussion of AWS should include less-than-lethal, non-lethal or non-kinetic weapons. Or whether there are particular types or classes of weapons that would be deemed non-problematic, such as countermeasure weapons. In other forums, the UK has stated that countermeasure weapons, such as close-in weapons systems like Phalanx that acquire and engage targets without human involvement, are exceptions to its position that humans always make acquisition decisions.[38] The doctrine states that "the operation of UK weapons will always be under human control as an absolute guarantee of human oversight, authority and accountability. Whilst weapon systems may operate in automatic modes there is always a person involved in setting appropriate parameters."[39]

Government of the United States of America

> "A weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation."[1]

The definition offered by the United States Department of Defense's 2012 Directive 3000.09 appears to be a functionalist approach to defining AWS, like that of the ICRC. However, because the US definition is embedded in a Department of Defense (DoD) policy, its purpose is ground discussions internal to US policy relating to the development and use of autonomous and semi-autonomous systems. As such, the definition needs to be considered with several other elements of the policy directive, in particular to demonstrate how the US's functional account is qualitatively different than the ICRC.

First, the definition ought to be taken in tandem with the definition of "semi-autonomous weapon systems". This is important because it shows where the primary distinction lies between the two classes of systems. A semi-autonomous weapon system is:

[38] Article 36, 2016, "The United Kingdom and Lethal Autonomous Weapons Systems: Analysis of UK Government Policy Statements on Lethal Autonomous Weapons Systems", http://www.article36.org/wp-content/uploads/2016/04/UK-and-LAWS.pdf.

[39] United Kingdom Ministry of Defense, 2017, *Unmanned Aircraft Systems,* Joint Doctrine Publication 0-30.2, p. 43, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/640299/20170706_JDP_0-30.2_final_CM_web.pdf.

> A weapon system that, once activated, is intended to only engage individual targets or specific target groups that have been selected by a human operator. This includes: (a) semi-autonomous weapon systems that employ autonomy for engagement related functions, including, but not limited to, acquiring, tracking, and identifying potential targets; cueing potential targets to human operators; prioritizing selected targets; timing of when to fire; or providing terminal guidance to home in on selected targets, provided that human control is retained over the decision to select individual targets and specific target groups for engagement. (b) "Fire and forget" or lock-on-after launch homing munitions that rely on TTP (tactics, techniques and procedures) to maximize the probability that only the targets within the seeker's acquisition basket when the seeker activates are those individual targets or specific target groups that have been selected by a human operator.[40]

Second, both of these definitions ground the essential policy directive that all "autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force" (4.a). What this judgment looks like, or what sorts of actions or indicators of situational awareness are deemed "appropriate", are left open.

The strength of the US's approach lies in its comprehensiveness and its attention to detail, both in terms of technical accuracy as well as in robust accounting of the types of design, development, acquisition, testing, evaluation and training that would be required throughout the DoD with regards to autonomous and semi-autonomous systems. This is the most detailed, publicly available policy in the world, and therefore there is more nuance to the approach contained within the policy as a whole.

The US places the focus on what constitutes autonomy at the level of *decision* rather than on the presence or absence of a particular technology. It is that the weapon system decides which target to attack; target "selection," therefore, is "the determination that an individual target or a specific group of targets is to be engaged."[41] The technologies that any particular system possesses to allow it to make this decision are mission and/or task specific, and so are not needed in the definition.

Moreover, as the Directive notes, autonomy in engagement-related functions does not mean the system is an "autonomous weapon system". Autonomy resides wherever the choice of target *and* employment of effects against that target is. If the choice resides with the operator, then a system is semi-autonomous. If the choice resides within the weapon system, after activation, then it is autonomous.

Unlike the ICRC's "critical functions" approach, the US does not look to the presence or absence of autonomy in particular engagement-related functions. While it is certainly true that the capability to make a "decision" entails that a weapon system will possess various functions that enable it to carry out this task, the US does not define autonomy purely by the presence or absence of those capabilities. In short, the US is not about "critical functions" in terms of listing various capabilities; it is rather about the decision-making process. One benefit of approaching the problem in this way is that as technological capabilities change with time, the definition does not require change.

However, the US approach raises several unanswered questions. First, how one describes the decision-making process and the time at which "selection" occurs is crucial. For example, because target-selection can often occur before the launch of any weapon system, there must be clear guidelines for commanders and operators for "selection" of targets before, and after, activation of

[40] Ibid.

[41] Ibid, p. 15.

a system to ensure that they comply with the different policy constraints within the DoD. There could be unintended or unforeseen erosion of decision-making authorities, for instance. For systems deployed for long periods of time, there may be instances where “selection” begins to occur without the commander’s knowledge or intent.

Second, this definition may have unintended consequences where commanders or operators begin to expand target selection to larger areas or classes of objects than originally intended by the policy. Instead of discussing individual targets or target groups, there may be incentives to demarcate whole areas as “targets” or target classes that are overly generalized (such as “vehicles” or “all military aged men”). This would keep the “decision” with the commander or operator, and thus categorize the weapon system as “semi-autonomous” though in use it would appear to be “autonomous” in its operation.

Additionally, the Directive is silent on whether or not selection occurs by default if a target recognition software has more than one type of target object in its library. For example, if a weapon system is able to identify M1 tanks, Apache Helicopters, and Patriot Missile batteries, would its deployment to a particular area to “hunt” for any of these objects and attack them constitute “selecting” or would the decision (select and attack) occur at the time of launch because the commander wanted to destroy one or all of those objects? If it is the latter, rarely will there be autonomous systems because commanders usually have precise targeting objectives.

Because the US definition is not directed towards particular kinds of technologies, we do not need to consider such questions as whether the system is: automatic, autonomous or multi-modal; is offensive or defensive; is anti-personnel or anti-materiel; or whether there is agreement on the “critical functions”. The primary consideration is whether the decision-making process to select and engage is handed over to the weapon system.

The biggest risk is that this will not be a bright line, and that in cases with long-deployed systems with little to no communications, systems with learning capabilities, or with systems that are systems of systems with individual semi-autonomous weapons systems acting together could inadvertently take over the decision-making process that the commander or operator believes she still has.

The US Directive explicitly excludes non-kinetic (cyber operations) systems from the policy.

5. Conclusions

There is a considerable and growing literature on all of the topics contained within this primer. As indicated in this primer, the range of concerns have been well described over the past three years of informal meetings of experts as well as by non-governmental experts. The definitional approaches have also been well described. Deeper understanding of the characteristics described in section II, as well as using more precise terminology in the CCW discussions, will contribute to more focused discussions going forward.

The 2017 GGE may make a recommendation to the CCW Meeting of High Contracting Parties to renew the mandate of the GGE in 2018. This makes sense—particularly since the 2017 meetings were cut short due to the financial situation of the Convention and that there is growing pressure—whether from civil society or industry groups—on the international community to take action on autonomous weapon systems. This is all the more urgent as the underlying technologies are advancing at a pace at which it is difficult for the international policy discussion to keep up.

Going forward, governments that are interested in making progress on addressing the issue of autonomy in weapon systems will need to decide the most productive way to do so in the limited time that is accorded to this activity within the disarmament calendar. One concrete approach would be for the High Contracting Parties to be more specific in the sequencing of activities within the GGE’s mandate in order to use the limited time in the most effective way.

UNIDIR

The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches

a primer

Agreeing on a working definition of LAWS will be a challenging endeavour, as there are several working definitions already in circulation, and some stakeholders have already stated a preferred policy response. Moreover, each proposed definition attends to a particular set of concerns and characteristics, while omitting others.

One's position on both an appropriate definition and an adequate policy response ultimately depends on what one is concerned about. Different definitions will attend to different sets of concerns, as well as privilege different sets of characteristics.

The objective of this primer is to consolidate and give an overview of both concerns and characteristics and illustrate how different definitional approaches attend to these.

UNIDIR RESOURCES