3rd International Conference on Public Policy (ICPP3)
June 28-30, 2017 – Singapore

Panel TO9P11 Session 2
The Governance of Innovative Technologies

Title of the paper

How to govern risks and uncertainties inherent in lethal autonomous weapon systems? Key legal challenges

Author
Mikolaj Firlej, Oxford University Faculty of Law, United Kingdom, mikolaj.firlej@law.ox.ac.uk

Date of presentation
Thursday, June 29th 13:30 to 15:30 (CJK 1 – 1)

KEYWORDS

Lethal autonomous weapon systems, international law, human control, autonomy

ABSTRACT

The lethal autonomous weapon systems (LAWS) are typically defined as weapons designed to select and attack targets without direct human control. It is said that a further development of these weapons will open for the first time the possibility of removing the human operator from the battlefield.

In my paper, I explore how to govern risks and uncertainties related to LAWS. First, I present what are the key characteristics of LAWS and how LAWS are different to currently existing weapons with a particular attention towards the independence from human operator. Then, I explore what are the key legal principles currently applicable to LAWS, in particular in the context of legal and moral significance of human judgment and attribution of responsibility. In the last part, I apply and critically evaluate my working definition of LAWS with reference to the key legal principles to determine the legal status of LAWS in the context of international law.

NOTE: The following piece is an excerpt from a much longer forthcoming essay-in-progress

INTRODUCTION

The rapid development of unmanned vehicles and their increasing usage in the military operations[1] only intensify technological development towards greater autonomy in the context of weapons systems. Although 'killer robots' – in a similar fashion to real-life Terminator Robot – is still a science fiction scenario, the incremental development of technology has sparked a popular and academic debate on the future of autonomous weapons, and the future of armed conflict in general.

The current debate is largely framed by fear that autonomous weapons could represent a new, dangerous category of weapons fundamentally distinct from the weapons of today. In spite of the fact that countries to a large extent publicly distance themselves from the development and potential deployment of so-called fully autonomous weapons[2], there is growing speculation within military and policy circles, especially in the US, that the future of armed conflict is likely to include lethal autonomous weapon systems. In 2013 leadership in the U.S. Navy and Department of Defence (DoD) identified autonomy in unmanned systems as a "high priority"[3], whilst in March 2016, the Ministries of Foreign Affairs and Defence of the Netherlands expressed their belief that "if the Dutch armed forces are to remain technologically advanced, autonomous weapons will have a role to play, now and in the future".[4] This

[1] Only in recent years unmanned aerials have been used in military operations in countries such as Pakistan, Yemen and Somalia.
[2] *Country Statements*, Convention on Certain Conventional Weapons Meeting of Experts on Lethal Autonomous Weapons Systems, Conference on Understanding Different Types of Risk, 2016 http://www.unog.ch/80256EE600585943/(httpPages)/37D51189AC4FB6E1C1257F4D004CAFB2?OpenDocument
[3] U.S. Department of Defense, *Unmanned Systems Integrated Roadmap: FY2013-2038*, 2013, p. 67.
[4] Government of the Netherlands, *Government Response to AIV/CAVV Advisory Report no. 97, Autonomous Weapon Systems: The Need for Meaningful Human Control*, 2016.

perception appears to be widely shared in the international community as over 40 countries working currently on robotic weapons[5]. The question of the role of autonomous systems in military operations becomes a matter of significance importance.

Fortunately, the discussion has been recently intensified mainly due to the controversial report launched by a 'Campaign to Stop Killer Robots', a group of activists from 54 non-governmental organisations. The report prompted state parties to the United Nations Convention on Certain Conventional Weapons (CCW) to held discussion on autonomous weapons. As a result of the last Review Conference (April 2016) of the High Contracting Parties to the UN Convention on Prohibitions or Restrictions on the CCW, Parties agreed on the establishment an open-ended Group of Governmental Experts (GGE). The GGE will explore and agree on possible recommendations on options related to the governance of autonomous weapons, in the context of the objectives and purposes of the Convention. At the conference, fourteen States have publically called for a pre-emptive ban on autonomous weapons.[6] Other States have taken an active role in the debate, including notably the United States, United Kingdom, China, Russia, India, Pakistan, Canada, France, Australia, the Netherlands, Belgium, and Germany.[7] In a similar manner, the UN Special Reporter on Extrajudicial, Arbitrary and Summary Executions, Christof Heyns produce a report which states that autonomous weapons requires not a ban but a moratorium of their development.[8] The moratorium is very timely as the debate has garnered a considerable attention of wide range of actors, including policy-makers, social activists, lawyers and philosophers. The promise of this debate is to clarify and structure the number of problems and challenges related to autonomous weapons in order to

[5] Sir Roger Carr, *Comments on World Economic Forum in Davos: "What if Robots go to war?"*, 2016.
[6] The following countries have publically endorsed a ban: Algeria, Bolivia, Chile, Costa Rica, Cuba, Ecuador, Egypt, Ghana, Holy See, Mexico, Nicaragua, Pakistan, State of Palestine, and Zimbabwe.
[7] *Country Statements*, Convention on Certain Conventional Weapons Meeting of Experts on Lethal Autonomous Weapons Systems, Conference on Understanding Different Types of Risk, 2016. http://www.unog.ch/80256EE600585943/(httpPages)/37D51189AC4FB6E1C1257F4D004CAFB2?OpenDocument
[8] Christof Heyns, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, 2013.

determine whether a ban or additional regulation is necessary.

Drawing from this promise, the topic of this article is to investigate whether existing international law of armed conflict sufficiently govern autonomous weapons. In particular, I am investigating four key principles of the international humanitarian law: military necessity, distinction, proportionality, and humanity.[9] These principles are derived from treaties such as the Hague Convention of 1907, the Geneva Conventions of 1949, and the 1977 Additional Protocols to the Geneva Conventions. The principles fit under a weapons' review process framework that is a requirement of law in order to field a new weapon.[10]

My argument asserts that although LAWS pose serious threat to international legal system, the current 'weak version' of autonomous weapons complies with key principles of international law. However, future developments of autonomous technologies (i.e. 'strong version of LAWS'), including advances in so-called general artificial intelligence, will require additional considerations beyond existing regulatory measures. The overall argument proceeds as follows: in order to appraise the conclusion, one must have a clear understanding of the notion of autonomous weapons. I conducted the clarification in two stages. First, distinction is being made by real and nominal definitions. Real definitions provide one's with the concrete thing or things denoted by defined object, whilst the nominal definition gives the meaning and use of defined object. I turned my attention into nominal definition, i.e. investigation of the existing meaning of autonomous weapon, as real definition does not seem to provide a sufficient explanation of LAWS. Second, taking into account the meaning currently assigned to LAWS, I specified three conditions which should be met in order to be considered as an autonomous weapon, i.e. a weapon should (1) posses first and second-order capacities; (2) meet substantive and procedural requirements of second-order capacities; and (3) posses a characteristic of

[9] Gary Marchant et al., *International Governance of Autonomous Military Robots*, 12 Colum. Sci. & Tech. L. Rev. 272 (2011)
[10] International Committee of the Red Cross Geneva, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977*, 2006.

'independence'. Thus for a weapon to have the second-order capacity means 'having the ability to critically evaluate the process of selecting and engaging target by considering reasons for their acceptance and for the acceptance of alternative actions, and to change them if they do not pass critical scrutiny'. In short, that means in practise of military operations that a particular weapon is capable of either changing a target previously picked out by a human operator or selecting its own target based on some information processing about this particular target. The second-order capacities have two components: (1) formal and (2) substantial. A formal component is just the ability of a weapon to conduct independent evaluation of targets and then engage with them accordingly. A substantial component is related to specific decisions which an object (weapon system, robot) is able to make. For example, such robot would be able to individually assess factors such as acceptable threshold of collateral damage, value of individual life and other factors which may potentially determine robots' decision which target to select and engage with. Weapons without substantial component are selecting targets and making relevant adjustments only in such a way and to the particular extent manufacturers programmed their goals. They do not learn on their own how to act in similar operational situations and they do not reveal any complex moral preferences, for example whether to eliminate an object or save its life. Such weapons are conducting only relatively simple evaluations how to be more effective under pre-determined categories. On the contrary, a weapon with substantive dimension of the second-order capacities will ultimately be similar to human beings. It is thus because such object would be able to make specific decisions with regard to specific situations, learning on their own, adjusting decisions and thus revealing certain complex action-preferences. It will be based on deep or general artificial intelligence and as such capable of understanding higher level intent and direction. Lastly, an autonomous weapon should be independent from human operator. As it seems, a weapon is doomed to be independent when there is either (1) no human involvement in the process of selecting and

engaging with targets or (2) there is an initial human involvement in framing the selection and engagement with target but ultimately the decision is delegated to a weapon.

Taking into account above categories, I differentiated between strong and weak LAWS. Strong LAWS are truly autonomous weapons which meet all three requirements, including substantial component. Such weapons do not exist yet. Weak LAWS should also meet all three requirements excluding substantial component. It is argued then that currently such weapons do exist and examples including The Long Range Anti-Ship Missile AGM-158C LRASM and IAI Harpoon. Strong and weak LAWS are differentiated from so-called independent weapons which are independent from human operator but are not able to critically evaluate target among the given selection of objects. Broad category of independent weapons is ranging from sophisticated "fire-and-forget systems" to simple landmines.

It is then argued that although strong version of LAWS may indeed challenge existing international regulatory regimes and should require additional regulation, the weak version is also incompatible with key principles of international law. Drawing from this analysis, it is argued also that any additional regulations for today's 'weak sense of LAWS' should be crafted with an eye toward tomorrow's 'strong sense of LAWS'.

In terms of methodology, the argument is based on both analytical analysis and comparative method. Analytical tools such as logical analysis of semantic terms as well as doctrinal legal analysis help to provide more clarity in this notoriously unclear notion of autonomy. Comparative method gives a better understanding where autonomous weapons should be placed within the wider framework of weapons' system. Drawing from this, I investigated how similar weapons have been regulated in the past, with a special emphasis on potentially suitable regulatory frameworks for LAWS. As there are some examples of pre-emptive bans on certain weapons, I explored whether those treaties provide a useful account for LAWS. Specifically, I explored a pre-emptive ban on blinding lasers stipulated in the Protocol

on Blinding Laser Weapons as well as the Ottawa Treaty which formally banned landmines in 1997. In the latter case, several states including US, China, Russia and India declined to sign the treaty, invoking military necessity. Nations that refused to sign the Treaty have generally complied with the more modest regulations of Amended Protocol. In a similar pattern, several states, invoking claims of military necessity, have declined to sign the Oslo Convention of 2008, which banned cluster weapons. These examples helps to provide an account whether LAWS should also be subject of additional regulation. With regard to potentially suitable framework for regulating LAWS, I investigated the Amended Protocol for landmines. The Amended Protocol on landmines focuses on geographic and spatial criteria for determining whether the deployment of mines is permissible. The Protocol makes clear that the definition of 'indiscriminate' use, a concept from the Geneva Conventions, applies to how landmines are placed. The Protocol requires additional protections in order to place certain weapons in areas containing a concentration of civilians and no active military engagement.

The structure of this paper is following: In the first chapter, I introduce a working definition of LAWS and briefly outline technological development in the field of military systems. In the second chapter, I will identify key legal principles applicable to LAWS deduced from international law of armed conflict. In the third part, I will apply and critically evaluate a working definition with reference to key legal principles to determine whether LAWS comply – or indeed whether they could comply with the reference to LAWS in strong sense – with existing international law. In the final part of the paper, the legal status of weak LAWS and strong LAWS is summarised.

Although in recent years many authors attempted to confront with the legal analysis of autonomous weapons[11], only few of them provided clarification with regard to the concept of

[11] See, among others, Alex Leveringhaus, *Ethics and Autonomous Weapons*, 2016; Dustin Lewis, Gabriella Blum, and Naz Modirzadeh, *War-Algorithm Accountability*, 2016; Michael W. Meier, *Lethal Autonomous Weapon Systems: Conducting a Comprehensive Weapons Review*, 2016.

'autonomy' in weapons system. Hence, I consider my contribution to the subject as twofold: First, by offering such clarification on the basis of the autonomous weapons' usage and meaning in the context of today's military operations; and secondly, by examining the extent to which current and future developments in the domain of autonomous weapons do not comply with the currently existing key principles of international law.

I. AUTONOMY IN WEAPON SYSTEMS

1. Why We need the Definition of Autonomous Weapons?

The purpose of this chapter is to delineate the scope of the concept of lethal autonomous weapon systems by providing maximally robust and comprehensive definition.

My decision to set out a working definition first rather than to take a one for granted is motivated by three main factors. Firstly, currently there is no internationally agreed-upon definition of what constitutes an "autonomous weapon," making communication on the topic very difficult[12]. Academics, respective national and international government organisations, including The United States Department of Defence, the United Nations and NGOs all use various definitions, and no standard is universally embraced.[13] Moreover, social activists campaigning to ban LAWS have yet to put forward a clear definition of their own or even clarify what, precisely, they are advocating should be banned.[14] Secondly, this lack of clarity in terminology is further amplified by the fact that some are calling for autonomous weapons to be regulated or banned even before consensus exists on how to define the category and thus understand better their legal status.[15] Third reason, LAWS are analytically and practically not

[12] Paul Scharre, "The Opportunity and Challenge of Autonomous Systems" in Andrew Williams, Paul Scharre (eds.), *Autonomous Systems. Issues for Defence Policymakers*, 2015
[13] The most popular definition is derived from US DOD Directive 3000.09 (US Department of Defence, *Directive 3000.09. Autonomy in Weapon Systems*, 2012): "LAWS is a weapon system that, once activated, can select and engage targets without further intervention by a human operator".
[14] Human Rights Watch, *Losing Humanity. The Case Against Killer Robots*, 2012.
[15] Paul Scharre and Michael C. Horowitz, *An Introduction to Autonomy in Weapon Systems*, 2015, p. 3.

easy to define, i.e. there are weapons that are autonomous to various degrees, and a definition still to be determined has to address the question how broad to draw the range of weapons included in the definition.[16]

This chapter aims to provide more clarity into discussed subject and determine whether one should view LAWS solely as weapons of the future or as present objects.

1.1.Definitional Limitations

The process of defining the objects is a challenging one as there are plethora of different types of definitions[17] and as such there is no accepted formula how to conduct a formal definition process, i.e. formal steps how to define an object and what are the conditions which every formal definition should met.[18] Thus, I focused my attention on already developed tools, including but not limited to the distinction between real and nominal definition as well as semantic analysis of terms, such as autonomous, automatic, automated.

Drawing from this caveat, firstly I think it should be noted that the selection of abovementioned tools has been limited to only few of them and there are other ways how to analyse the notion of autonomous weapons.

Secondly, my definition(s) reflects current stage of technological development of autonomous systems, however - as the technology is moving into greater autonomy – one cannot claim that what we currently identify as autonomous or self-learning systems will be similarly denoted in the future. Clearly, we do not know how exactly the development of intelligent systems will play out in terms of our understanding of autonomy, especially in the long-term.

---

[16] *Id*., p. 5.
[17] Guy Longeworth, "Definitions. Uses and Varieties of" in Alex Barber, Robert Stainton (eds.) *Concise Encyclopedia of Philosophy of Language and Liguistics*, 2010, pp. 138-142.
[18] A mainstream theory of definitions called sometimes the standard theory places the eliminability and conservativeness requirements on definitions. See more: Nueal Belnap, 'On Rigorous Definitions', *Philosophical Studies 72(2- 3)*, 1993, pp. 115–146.

Lastly, although my definition(s) of LAWS is provided in abstract terms, hence applicable irrespectively to any factual stance, one should note that it is heavily based on publicly available information on current technological development with regard to the autonomy in military systems. Such information may only partially reveal the real development of autonomous weapons.

On a linguistic note, for the purpose of this thesis, I use terms 'lethal autonomous systems', 'lethal autonomous weapon systems', 'lethal autonomous robots', 'lethal self-learning systems' and 'lethal super-intelligent systems' interchangeably, although they may not always be qualified as synonyms[19].

1.2.Nominal and Actual Definition of LAWS

Although, there is no accepted formula how to conduct formal definition process, I found it useful the distinction between nominal and real definition to investigate the semantic content of autonomous weapons.

According to John Lock, nominal essence is the abstract [Idea] to which the name is annexed, whilst real essence presents a defined object as a compound of other propositions or properties or relations.[20] In short, real (essentialist) definitions provides one's with the concrete thing or things denoted by defined object, whilst the nominal definition gives the meaning and use of defined object.

Taking Lock's framework into account, a real definition of LAWS should be based on its integral parts which constitute the machine capable of making lethal decisions. From essentialist perspective, computer algorithms are key ingredients of what most people associate with LAWS. An algorithm is usually defined in two ways: (1) as input/output relationship or as

[19] Dustin Lewis, Gabriella Blum, and Naz Modirzadeh, *War-Algorithm Accountability*, 2016, p. 6.
[20] John Locke, *An Essay Concerning Human Understanding*, Book III, Chapter VI, 1996.

(2) a tool for solving computational problem. According to the former dimension, an algorithm is "any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output."[21] According to the latter dimension, an algorithm describes a specific computational procedure for achieving a problem described in input/output relationship.[22] In the context of autonomous devices, the second definition comes in handy because machines are programmed in such a way to solve important military problems, such as the improvement of selecting and targeting capabilities. Thus, in the context of military, such algorithms are often called 'war algorithms'[23]. A "war algorithm" is any algorithm that is expressed in computer code, that is effectuated through a constructed system, and that is capable of operating in relation to armed conflict. Those systems include self-learning architectures that today offers a promise of "replacing" human judgment with algorithmically-derived "choices."[24]

Although an investigation into all essential parts of autonomous weapons, including their algorithms, may prove to be an insightful intellectual journey, I am not convinced that real/essentialist definition will provide much explanatory progress. The reason is that under the real definition one should investigates all relevant parts of autonomous weapons which seems to be an unrealistic *tour de force* given that we deal with sophisticated and interrelated computer algorithms. Nevertheless, Lewis et al. envisaged abovementioned approach by narrowing down war algorithms to only those that fulfil three conditions, namely: algorithms (1) that are expressed in computer code; (2) that are effectuated through a constructed system; and (3) that are capable of operating in relation to armed conflict.[25] However, this essentialist approach still does not seem to provide a significant explanatory understanding. Firstly, the notion of

[21] Dustin Lewis, Gabriella Blum, and Naz Modirzadeh, *War-Algorithm Accountability*, 2016, pp. 15-16.
[22] *Id*., p. 16.
[23] War-algorithms approach to LAWS has been identified by the Harvard Law School Program on International Law and Armed Conflict (PILANC). See: https://pilac.law.harvard.edu/aws/ .
[24] Dustin Lewis, Gabriella Blum, and Naz Modirzadeh, *War-Algorithm Accountability*, 2016, p. 7.
[25] *Id.*, p. 16.

computer codes that are capable of operating in armed conflict is very broad and does not specify what makes LAWS unique compared to other currently existing weapons. In another words, the real definition does not stipulate the difference between terms such as 'automation, 'autonomy' or 'independence' when it comes to weapons. As it seems, all of these terms carry a different meaning which should be elaborated. Secondly, the potential application of such definition in terms of legal interpretation gives little promise of success as law in general, and international law specifically, does not refer to algorithms but to people and states, and their relations with weapons. Interestingly, although Lewis et al. presented algorithm-based definition of LAWS in depth, when it comes to legal interpretation the Authors decided to use definitions derived from States' positions and their understanding of autonomous weapons.

Lastly and more broadly, real definitions rarely provide a useful account as the terms used in the definiens are hardly ever easier to understand than the definiendum. In LAWS example, autonomous weapons are difficult to define and war-algorithms seems to be equally conceptually difficult to grasp.

Hence, I suggest to focus on meaning and *use* of LAWS rather then on what constitute their real essence. In terms of explaining the meaning of any given object, one may assign the existing meaning or projected meaning of a specific object. For the purpose of this thesis, I shall denote a definition based on existing meaning as actual, whilst a definition based on projected meaning as notional.[26] As we generally speak about the future when it comes to LAWS and the meaning of what we associate as autonomous or general intelligence is changing with respect to technological development, I decided to focus only on analytical definition, i.e. on the current meaning and use of LAWS today rather than projecting future meaning and use of such weapons.

[26] See similar distinction in Jacek Jadacki, *Spór o granice języka*, 2002, p. 191. Jadacki is using the distinction between 'analytical' and 'synthetical' which I found slightly misleading from the linguistic perspective.

2. *The Concept of Autonomy: First- and Second-order Capacities*

In order to capture the meaning of autonomous weapons today, one should investigate what is autonomy and what autonomy means in the context of weapons in today's world. This sub-chapter though is not to present the history of autonomy and variations of this term but to provide useful distinctions for understanding of its meaning in the context of weapons systems.

As it seems, the key problem is lack of standardised definition of autonomous weapons, and thus experts are divided whether LAWS actually exist or are yet to be developed. On the one hand, "there is a nearly universal consensus, among both ban advocates and sceptics, that autonomous weapon systems do not yet exist."[27] On the other hand, even if one accepts their existence, one may have very different view of what constitutes an autonomous device. As a striking example, the term 'autonomous robot' denotes such dissimilar images as a household Roomba and the sci-fi Terminator, to name few.[28]

Probably the most widely adopted definition was formulated by U.S. Department of Defense policy directive which states that "autonomous weapon system" is a one "that, once activated, can select and engage targets without further intervention by a human operator."[29] The definition encompasses also so-called human-supervised autonomous weapon system: "An autonomous weapon system that is designed to provide human operators with the ability to intervene and terminate engagements, including in the event of a weapon system failure, before unacceptable levels of damage occur."[30] The directive also stipulates so-called semi-autonomous weapon system defined as: "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".[31] Directive establishes that, as a matter of policy, "autonomous and semi-

[27] Rebecca Crootof, *The Killer Robots are here: legal and policy implications*, 2015, p. 1863.
[28] Paul Scharre and Michael Horowitz, *An Introduction to Autonomy in Weapon Systems*, 2015, p. 4.
[29] US Department of Defense, *Directive 3000.09: Autonomy in Weapon Systems*, 2012, p. 13.
[30] *Id*., p. 14.
[31] *Id*.

autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force".[32]

In a rather striking contrast, France articulated the following considerations with respect to LAWS definition: "Lethal autonomous weapons systems are fully autonomous systems. Remotely operated weapons systems and supervised weapons systems should not be regarded as LAWS since a human operator remains involved, in particular during the targeting and firing phases. Existing automatic systems are not LAWS either. 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."[33]

Compared to most other states that have put forward working definitions, France articulates a narrow definition of what constitutes a lethal autonomous weapons system. Most striking, perhaps, is the condition that there be "a total absence of human supervision, meaning there is absolutely no link (communication or control) with the military chain of command." Thus, according to the French definition, LAWS are future weapon systems: they do not currently exist. According to the US definition, DoD stipulates broader conditions for weapons to be considered as LAWS, including a presence of human supervision. Taking this definition into account, LAWS already exist, not exclusively in the US. Referring to Sharre and Horowitz research, currently around 30 countries have defensive systems with human-supervised autonomous modes that are used to defend military bases and vehicles from short-

[32] *Id*.
[33] The Government of France, *Non-paper. Characterization of LAWS for the Convention on Certain Conventional Weapons Meeting of experts on Lethal Autonomous Weapons Systems in Geneva*, 2016.

warning attacks, where the time of engagement would be too short for a human to respond.[34] Examples include close-in weapon systems such as the Goalkeeper, AK-360 and Phalanx.[35]

Although French definition puts strict requirements for a weapon to be classified as autonomous, the highest bar has been set arguably by the U.K. Ministry of Defence. The U.K. defines LAWS as a weapon system that "will, in effect, be self-aware"[36] as it is "capable of understanding higher level intent and direction. (…) 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".[37] In accordance to U.K. definition LAWS are related to very sophisticated artificial intelligence algorithms, far beyond the technological realm of current developments.

Clearly, the purported existence of LAWS is not just a matter of linguistic unambiguity. At the heart of linguistic difficulties is the question to what extent one should consider the usage of LAWS as a new category for qualitatively different weapons. In another words, to what degree currently existing weapons present as unique compared to weapons in the past, such as, for example, landmines.

In order to answer this question one may begin with etymological considerations. Traditionally the concept of autonomy has been applied to personal or moral autonomy, although not exclusively. In general, the term "autonomous" derives from the Greek words *autos*, meaning "self," and *nomos*, meaning "rule." The etymology of the word directly informs the common meaning of the term, often understood as something that is self-governing. Usually, self-governing is related to the agent's freedom to choose her desires, preferences and goals. A more formal approach to autonomy can be elaborated in a following way: 'having

[34] Paul Scharre and Michael Horowitz, *An Introduction to Autonomy in Weapon Systems*, 2015, p. 3.
[35] Michael Horowitz, *Why words matter: The real world consequences of defining autonomous weapons systems*, 2016, p. 88.
[36] U.K. Ministry of Defence, *Joint Doctrine Note 2/11: The UK Approach to Unmanned Aircraft Systems*, 2011, pp. 2-3.
[37] *Id*.

freedom to choose one's goals' means ''having the second-order capacity to reflect upon and distance oneself from one's first-order capacities, i.e., action-guiding goals, to evaluate them in a critical way by considering reasons for their acceptance and for the acceptance of alternative goals, and to change them accordingly if they do not pass critical scrutiny'.[38] One shall not be called an autonomous agent if one does not express second-order capacities. This is so because the agent's first-order, i.e. action-guiding goals can be really called hers only if she evaluates them and ultimately accepts these goals, i.e., if she has second-order goals to have and to realise her first-order goals.

The presented concept of second-order capacities has two dimensions: substantive and procedural dimension.[39] The procedural dimension only requires that an agent should be able to critically evaluate her first-order goals, and, should these goals not pass the critical evaluation, to change them accordingly. This condition does not stipulate for an agent to choose a specific type of goals accordingly to specific situation and certain principles of behaviour. The procedural dimension only put emphasis on formal process of reviewing first-order capacities. However, the evaluation procedure does not include any requirements about the content of the goals in virtue of which one is considered autonomous. Hence, an object which possess only procedural dimension can be just as well an egoistic utilitarian interested only in maximising a certain variable (i.e. the total amount of damage) as an altruist treating others' needs as one's own needs. A truly autonomous object should not be understood as consistent with certain constrained evaluation of goals no matter how an object came to choose such goals. Accordingly, the substantive dimension requires that an agent should choose a specific type of goals relative to specific situation, i.e. revealing certain moral preferences, realising moral

[38] The concept of higher-order volitions is typically associated with Harry Frankfurt. See Harry Frankfurt, *Freedom of the Will and the Concept of a Person*, 1971. See also Gerald Dworkin, *The Theory and Practice of Autonomy*, 2001.

[39] Some writers have insisted that the autonomous object must enjoy substantive (material) as well as procedural (formal) component, especially in the context of personal autonomy, i.e. Mal Oshana, *Personal Autonomy in Society*, 2006; Natalie Stoljar, *Autonomy and the Feminist Intuition*, 2000.

principles, rational course of action, etc. A different explanation of substantial component is based on the importance of the individual history of the agent as an element of her autonomy.[40] In harmony with this view, the question of whether an object is autonomous at a time depends on the past history of processes by which an object came to be the way it presents today. Thus, an object which is able to take only single decision (as it is with many so-called autonomous weapons) does not have substantial component irrespectively whether it is able to evaluate critically such decision.

The concept of autonomy is closely related to the concept of freedom. As it seems, freedom at the level of realising one's goals should be distinguished from freedom understood as the lack of obstacles in realising one's goals.[41] The latter sense of freedom I call 'independence'. These two dimensions of freedom are essential for an object to be consider as autonomous in practice.

2.1. Autonomy in the context of weapon systems

Drawing from above considerations, to be a fully autonomous, a weapon should meet following requirements: (1) posses first and second-order capacities; (2) meet substantive and procedural requirements of second-order capacities; and (3) posses a characteristic of 'independence'. This chapter aims to evaluate currently existing weapons against these metrics.

As the concept of first and second order capacities is typically related to personal autonomy[42], one can reformulate these terms in a more formal language in the context of weapon systems. Thus, for a weapon to have the second-order capacity means 'having the ability to critically evaluate the process of selecting and engaging target by considering reasons for their acceptance and for the acceptance of alternative actions, and to change them if they do

[40] John Christman, *Autonomy and Personal History*, 1991.
[41] The concept reflects Isaiah Berlin's approach in *Two Concepts of Liberty*, 1958.
[42] Harry Frankfurt, *Freedom of the Will and the Concept of a Person*, 1971

not pass critical scrutiny'. In short, that means in practise of military operations that a particular weapon is capable of either changing a target previously picked out by a human operator or selecting its own target based on some information processing about this particular target.

In order to elaborate the characteristic of first and second-order capacities and substantive and procedural requirements four weapons are compared which are sometimes denoted altogether as 'autonomous'[43]: LRASM anti-ship missile, Brimston fire-and-forget missile, smart munitions and landmines.

The Long Range Anti-Ship Missile AGM-158C LRASM is one of the most advanced weapon publicly known.[44] Unlike most of current anti-ship missiles like U.S. Harpoon, the LRASM is capable of conducting autonomous targeting, relying on on-board targeting systems to independently acquire the target without the presence of prior, precision intelligence, or supporting services like Global Positioning Satellite navigation and data-links.[45] As is seems, LRASM and some other advanced weapons[46] possess limited second-order capacities as the weapon is able to self-select targets and make its own adjustments after is launched. LRASM is initially directed by pilots, but then halfway to its destination, it severed communication with its operators. The weapon itself decided which of selected targets to attack among the given pool.[47] The second-order capacities of such weapons lack substantive dimension as a weapon does not choose specific type of goals accordingly to specific situation. Rather, such weapon is able to review their own goals (selecting and engaging targets) in a limited fashion only in accordance with pre-programmed algorithms. In another words, a weapon can learn from itself to be more effective and better select appropriate targets in order to maximise damage but is

[43] Mary Ellen O'Connell, *Banning Autonomous Killing*, 2013, p. 6.
[44] Lockheed Martin Press Release, *Lockheed Martin Long Range Anti-Ship Missile Conducts Successful Jettison Flight Test from US Navy F/A-18E/F*, 2017. See http://www.lockheedmartin.co.uk/us/products/LRASM/mfc-lrasm-pressreleases.html
[45] *Id*.
[46] Another example is Israeli's IAI Harop (IAI Harpy 2), a type of loitering munition. The IAI Harop is able to operate fully autonomously, using its anti-radar homing system.
[47] See http://www.lockheedmartin.co.uk/us/products/LRASM/mfc-lrasm-pressreleases.html

not able to assess factors such as acceptable threshold of collateral damage, value of individual life and other factors which may potentially determine weapons' decision which target to select and engage with. Today's weapons are selecting targets and making relevant adjustments only in such a way and to the particular extent manufacturers programmed their goals. Thus, self-targeting weapons do not reveal any complex moral preferences but only strive to be more effective under pre-determined categories.

Less advanced weapons such as Brimstone "fire and forget' missiles can distinguish among tanks and cars and buses without human assistance but they are not capable to learn from their actions and evaluate their own goals. Therefore, such weapons should be classified as having only first-order capacities. They are able to select their own targets (i.e. realising their own goals) but lack critical evaluation, adjustment or improvement of their actions. Similarly, so-called smart munitions have only first-order capacities. Smart munitions are precise in their ability to locate a particular location in time and space. However, that location, either painted by a human being with a laser, or guided through coordinates and satellites, is set by a human. The human chooses that target (i.e. weapon's goals), not a weapon itself. A less advanced examples are landmines which are simple, automatic systems, unable to distinguish among the selection of potential targets.

In literature the distinctions between such weapons are often blurred but often follow a similar pattern. One of the most widespread division is between 'automatic', 'automated' and 'autonomous' weapon. The term "automatic" is used to refer to systems that have very simple, mechanical responses to environmental input, such as mines. The term "automated" is used to refer to more complex, rule-based systems, such as fire and forget systems. The word "autonomous" is reserved for machines that execute some kind of self-direction, self-learning or emergent behaviour that is not directly predictable from an inspection of its code. According to Noel Sharkey, an autonomous robot "is similar to an automatic robot except that it operates

in open and unstructured environments". The robot is still controlled by a program but now receives information from its sensors that enable it to adjust the speed and direction of its motors as specified by the program".[48] Although I found useful these distinctions, I am not convincing that 'open and unstructured environment' is something that ultimately differs autonomous weapon from other weapons, especially given how long-range weapons have been already successfully developed. Rather, an ability to evaluate their own goals in a given area is something that makes autonomous weapons unique.

The concept of 'independence', i.e. a lack of obstacles in realising one's goals in the context of weapon systems is recognised as lack of human involvement or human control over the performance of a weapon.[49] However, the concept of 'performing an action by machines' requires further elaboration. In military literature the term is often used as a synonym to 'selecting and engaging targets'. Drawing from this elaboration, a weapon is clearly dependent on human operator when the human decides to launch a weapon into an area where she believes enemy combatants exist. On the contrary, a weapon is regarded as independent if the human operator is not personally picking out the targets, but the weapon itself. The issue becomes more complicated when there is a mix of human and machine input throughout the process of selecting and engaging with a target. Michael Horowitz presents a following scenario: "Now imagine that instead of a human operator launching the missiles, a human operator believes there are enemy naval surface ships 1000 miles away, but is not sure if the ships are there. The human operator launches a robotic system programmed by algorithm to launch missiles at adversary ships if those ships are detected. The robotic system detects the ships and launches the missiles. (…) A human operator believed there were ships around that location and launched a weapons platform to attack those ships. The human is still doing the selection and engagement

[48] Noel Sharkey, *Automating Warfare: Lessons Learned from the Drones*, 2012.
[49] Defining 'independence' in the context of weapons is widely recognised as a lack of human involvement. See IHRC, *Killer Robots and the Concept of Meaningful Human Control Memorandum to Convention on Conventional Weapons (CCW) Delegates*, 2016

despite the uncertainty involved about the location of the ships. (…) The weapon platform is run by algorithm and responsible for the details of the engagement— finding, fixing, tracking, targeting, and terminally engaging the enemy ships."[50] Does the weapon should be classified as independent? Two arguments raised for this case. First, a weapon is dependent on a human only if a human can control or override a weapon's decision. If a particular weapon is *actively* searching for target and *actively* engaging with a target, then it is clear that a human does not possess control over its actions. Second, although a human operator selects and engages with targets initially, her involvement is only limited to general terms. In another words, a human controller specifies only general target and general engagement whereas details are to be determined by a weapon itself. Thus, a weapon possesses a wide operational power which is out of the control for human operator and largely influences the outcome of weapon's engagement. Thus, it is reasonable to claim that weapons described in example can be classified as independent.

A useful account for clarifying the level of human involvement over weapons provides Human Rights Watch Report, which divides robotic weapons into three categories: (1) Human-in-the-Loop Weapons, i.e. robots that can select targets and deliver force only with a human command; (2) Human-on-the-Loop Weapons, i.e. robots that can select targets and deliver force under the oversight of human operator who can override the robot's actions; (3) Human-out-the-Loop Weapons, i.e. robots that are capable of selecting targets and delivering force without any human input or interaction.[51] It has been widely recognised that machines that perform a function for some period of time, then stop and wait for human input before continuing, are often referred to as "semi-autonomous" or as having a "human in the loop."[52] On a related note, machines that can perform a function entirely on their own but have a human in a monitoring

[50] Michael Horowitz, *Why Words Matter: The Real World Consequences of Defining Autonomous Weapons Systems*, 2016, p. 88.
[51] Human Rights Watch, *Losing Humanity. The Case Against Killer Robots*, 2012.
[52] Paul Scharre and Michael Horowitz, *An Introduction to Autonomy in Weapon Systems*, 2015, p. 6.

role, with the ability to intervene if the machine fails or malfunctions, are often referred to as "human-supervised autonomous" or "human on the loop." Hence, machines that can perform a function entirely on their own with humans unable to intervene are referred to as "independent" or "human out of the loop." Taking into account Horowitz example, two further caveats should be added. First, a weapon shall be regarded as independent when there is either (1) no human involvement in the process of selecting and engaging with targets or (2) there is an initial human involvement in framing the selection and engagement with target but ultimately the decision is delegated to a weapon. Second, initial limited human involvement is clearly something different than 'human on the loop' concept. The former stipulates that there is no further human control after the initial selection and engagement process, whilst the latter assumes a constant human oversight irrespectively of the stage of selection and engagement process.

2.2.Two types of LAWS: strong and weak sense

As it seems, there is no existing weapon which satisfies all requirements for a truly autonomous object, i.e. (1) possesses first and second-order capacities; (2) meets substantive and procedural requirements of second-order capacities; and (3) possesses a characteristic of 'independence'. In particular, the ability for a weapon to meet the condition of substantive dimension of second-order capacities seems to be very far from the current technological development and it is unclear whether this is even feasible in the long-term future.[53] A weapon with substantive dimension of the second-order capacities will ultimately be similar to human beings.[54] It is thus because such object would be able to make specific decisions with regard to specific situations and thus revealing certain complex moral preferences. It will be close to what the U.K. definition of LAWS as a weapon that "will, in effect, be self-aware" and "capable of

[53] Robert Sparrow, "Robots and Respect: Assessing the Case Against Autonomous Weapon Systems" in *Ethics and International Affairs*, 30, no.1, 2016, p. 94.
[54] *Id*.

understanding higher level intent and direction." Although the purpose of this paper is not to elaborate whether substantive dimension of second-order capacities is equal to the characteristic of self-awareness, it seems that these terms are similar in a fashion that they both imply for a technology to enable human-like cognition. Number of experts in artificial intelligence strongly oppose such definition of LAWS exactly on the grounds that the current technology does not enable human like cognition and it is very unlikely to develop such technology in foreseeable future.[55]

However, there are weapons, such as LRASM, which are capable to express simple second-order capacities. The difference between LRASM and a truly autonomous object is that 'LRASM-type of weapons' are only able to satisfy procedural requirement of second-order capacities. LRASM does not make any complex decisions regarding specific situations, for example assessing whether to kill or not, what is the acceptable collateral damage and so on. The weapon is only able to make their own limited decisions regarding the selection of a target by using its onboard AI to locate a specific warship from among a fleet of enemy warships. A multi-mode seeker guided by the AI ensures the correct warship is hit in a specific area to maximize the probability of sinking the target. The scope of a target is determined by a human but the target identification is handed over to a weapon itself. Artificial intelligence system enables autonomous targeting and engaging with a target by using on-board targeting information to independently acquire a target without the need for prior human decision, Global Positioning Satellite navigation or data-links.

A truly autonomous weapon would be able to evaluate their own decisions in more sophisticated way, similar to human cognition processes. Assuming that weapons of future will use an advanced combination of facial and image recognition, as well as infra-red and radar systems. Assuming further, such weapons will not be constrained to air domain but will

[55] Noel Sharkey, *Automating Warfare: Lessons learned from the Drones*, 2012, pp. 2-3.

populate maritime and ground environments. A truly autonomous weapon then would be able to select their own target based on, say, Instagram picture and then will have an ability to find a relevant target wherever located and decide whether to take any action. A sequence of similar actions will probably reveal certain preferences and patterns in a similar fashion to humans "morality". This is the substantial component of second-order capacities that currently existing weapons do not have.

What makes LRASM similar to a truly autonomous weapons, is the ability to possess second-order capacities, i.e. capacity to evaluate before making final decisions. Although LRASM has a relatively simple procedure and potential target zone is predefined by a human operator, ultimately LRASM weapon is able to evaluate a decision and select the best possible target on their own. While the exact factors that constitutes the target selection algorithm are classified, it is likely a weighting of elements such as the target's size, location, radar signature, heat profile, or other elements that positively identify and allows to engage with the target.

Another similar feature between 'LRASM-type of weapons' and a truly autonomous weapons is their independence from human operator. 'LRASM-type of weapons' selects and engage targets on its own, without a human "in" or "on the loop". For example, weapons such as loitering munitions are not launched at a specific target. Rather, they are launched into a general area where they will loiter, flying a search pattern looking for targets within a general class, such as enemy radars, ships or tanks. Then, upon finding a target that meets its parameters, the weapon will fly into the target and destroy it. Thus, unlike for example a homing munition, the human operator launching the weapon does not know the specific target that is to be engaged, only that the weapon will engage targets of a particular class within a broad geographic area. The most notable example of such weapon is the IAI Harop (IAI Harpy 2). The Harop is part-UAV, part-missile development in which the entire aircraft becomes an attack weapon upon spotting a target of opportunity. It is a "hunter-killer" UCAV system that can

loiter in a given area, survey potential targets movements and hunt for critical targets. The drone can either operate fully autonomously, using its anti-radar homing system, or it can take a man-in-the-loop mode. If a target is not engaged, the drone will return and land itself back at base.

As it seems now, LRASM-type of weapons share some important characteristics with a truly autonomous weapons, i.e. second-order capacities and independence from human operator ("human out of loop). Hence, I decided to classified them as lethal autonomous weapon systems in a weak sense (LAWSw), whilst a hypothetical truly autonomous weapon with substantial component of second-order capacities will be denoted as lethal autonomous weapon system in a strong sense (LAWSs).

Besides LAWSw and LAWSs there are certain weapons which are fully independent from human operator but do not possess second-order capacities. Such weapons are for example "fire and forget systems", landmines and others. Those weapons however will not be subject of my legal analysis.

Below a table which summarize the main types of weapons, their characteristics and examples.

Type of weapon	**LAWS strong sense**	**LAWS weak sense**	**Independent weapons**
Features	1st and 2nd order capacities	1st and 2nd order capacities	-
	Substantial and Procedural dimension	Procedural dimension	-
	Out of the loop	Out of the loop	Out of the loop

Examples	Yet to be developed	IAI Harop, LRASM	Fire-and-forget weapons, i.e. Phalanx Close-in Weapon System, Landmines and others

The above definitions have been constructed based on meaning and use of autonomous functions in the today's context of weapon systems. Some other authors, notably Robert Sparrow, similarly identified what I refer to as 'LAWS in a weak sense'. Sparrow defined LAWSw as weapons "which are capable to being tasked with identifying possible targets and choosing which to attack without human oversight, and that is sufficiently complex which such that, even when it is functioning perfectly, there remains some uncertainty about which objects and/or persons it will attack and why".[56] These weapons are already developed and their legal status shall be a subject of the next chapters.

REFERENCES (selected)

Belnap, Nueal, (1993) 'On Rigorous Definitions', *Philosophical Studies 72(2- 3)*.

Berlin, Isaiah, *Two Concepts of Liberty*, 1958.

Carr, Roger, *Comments on World Economic Forum in Davos:"What if Robots go to war?"*, 2016.

Christman, John, *Autonomy and Personal History*, 1991.

[56] Robert Sparrow, "Robots and Respect: Assessing the Case Against Autonomous Weapon Systems" in *Ethics and International Affairs*, 30, no.1, 2016, p. 95.

Country Statements, Convention on Certain Conventional Weapons Meeting of Experts on Lethal Autonomous Weapons Systems, Conference on Understanding Different Types of Risk, 2016.

Crootof, Rebecca, *The Killer Robots are here: legal and policy implications*, 2015.

Dworkin, Gerald, *The Theory and Practice of Autonomy*, 2001.

Frankfurt, Harry, *Freedom of the Will and the Concept of a Person*, 1971.

Government of France, *Non-paper. Characterization of LAWS for the Convention on Certain Conventional Weapons Meeting of experts on Lethal Autonomous Weapons Systems in Geneva*, 2016.

Government of the Netherlands, *Government Response to AIV/CAVV Advisory Report no. 97, Autonomous Weapon Systems: The Need for Meaningful Human Control*, 2016.

Heyns, Christof, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, 2013.

Horowitz, Michael, *Why Words Matter: The Real World Consequences of Defining Autonomous Weapons Systems*, 2016.

Human Rights Watch, *Losing Humanity. The Case Against Killer Robots*, 2012.

International Committee of the Red Cross Geneva, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977*, 2006.

International Human Rights Commission, *Killer Robots and the Concept of Meaningful Human Control Memorandum to Convention on Conventional Weapons (CCW) Delegates*, 2016

Jadacki, Jacek, *Spór o granice języka*, 2002.

Leveringhaus, Alex, *Ethics and Autonomous Weapons*, 2016.

Lewis, Dustin, Blum, Gabriella and Modirzadeh, Naz, *War-Algorithm Accountability*, 2016.

Locke, John, *An Essay Concerning Human Understanding*, Book III, Chapter VI, 1996.

Lockheed Martin Press Release, *Lockheed Martin Long Range Anti-Ship Missile Conducts Successful Jettison Flight Test from US Navy F/A-18E/F*, 2017.

Longeworth, Guy, (2010) "Definitions. Uses and Varieties of" in Barber, Alex, Stainton, Robert (eds.) *Concise Encyclopedia of Philosophy of Language and Liguistics*.

Marchant, Gary et al. (2011), "International Governance of Autonomous Military Robots" in *12 Colum. Sci. & Tech. L. Rev*. 272.

Meier, Michael, *Lethal Autonomous Weapon Systems: Conducting a Comprehensive Weapons Review*, 2016.

O'Connell, Mary Ellen, *Banning Autonomous Killing*, 2013.

Oshana, Mal, *Personal Autonomy in Society*, 2006.

Scharre, Paul, Horowitz, Michael, *An Introduction to Autonomy in Weapon Systems*, 2015.

Scharre, Paul, (2015) "The Opportunity and Challenge of Autonomous Systems" in Williams, Andrew, Scharre, Paul (eds.), *Autonomous Systems. Issues for Defence Policymakers*.

Sharkey, Noel, *Automating Warfare: Lessons Learned from the Drones*, 2012.

Sparrow, Robert, (2016) "Robots and Respect: Assessing the Case Against Autonomous Weapon Systems" in *Ethics and International Affairs*, 30, no.1.

Stoljar, Natalie, *Autonomy and the Feminist Intuition*, 2000.

U.K. Ministry of Defence, *Joint Doctrine Note 2/11: The UK Approach to Unmanned Aircraft Systems*, 2011.

U.S. Department of Defense, *Unmanned Systems Integrated Roadmap: FY2013-2038*, 2013.

U.S. Department of Defence, *Directive 3000.09. Autonomy in Weapon Systems*, 2012.