Research Article Compendium

Peace and Conflict Research

Autonomy in Future Military and Security Technologies: Implications for Law, Peace, and Conflict

RICHARDSON INSTITUTE

The Richardson Institute, Lancaster University, UK

The Richardson Institute is the oldest Peace Studies centre in the UK and was established in 1959 in the spirit of the Quaker scientist, Lewis Fry Richardson. The Richardson Institute is an interdisciplinary forum for research on peace and conflict.

Based within the Department of Politics, Philosophy and Religion (PPR) at Lancaster University, the Richardson Institute brings together academics who are committed to undertaking contemporary research and to developing outreach activities.

Our vision is to be internationally recognised as a world-class partner of choice for universities, research centres and civil society stakeholders on research into peace and conflict and knowledge exchange activities.

For more information, please see: http://www.lancaster.ac.uk/fass/centres/richinst/index.htm

---

The Faculty of Law, University of Barcelona, Spain

The Faculty of Law is the oldest surviving institution at the University of Barcelona and, indeed, one of the most historic faculties in Catalonia.

It has provided a large number of higher education courses over the years, through the various law schools that have emerged from the Faculty and the diverse programs of study these have offered, as well as through courses delivered at the Faculty itself. These have traditionally included studies related to the public service and specific degrees in criminology and private investigation.

Based within the Faculty of Law, the Department of Criminal Law and Criminal Science, and Public International Law and International Relations of Politics, brings together academics who are committed to undertaking contemporary research and to developing outreach activities. The Department is formed by several academics who carry out high-quality specialized teaching, research and knowledge transfer activities in their fields of expertise.

Its main goal is to promote the development of high level scientific research under an interdisciplinary, international and prospective approach which achieves a contribution to social and economic progress, improving people's living conditions, defending the inherent dignity of all members of the human family and the right to life, and contributing to building a better and fairer society for everybody, everywhere, in security, peace and dignity.

For more information, please see: https://www.ub.edu/portal/web/law/.

Cover Photo: U.S. Navy X-47B Unmanned Combat Air System demonstrator aircraft aboard USS George H.W. Bush May 14, 2013. Photo taken by MC2 Timothy Walter. Used under public domain license. Photograph from defenseimagery.mil.

Contents

Introduction and structure of the report

It is undeniable that without science and engineering, contemporary warfare as we know it would not exist. History shows us how technological advancements have, over the years, been a catalyst to the development of security and defence weaponry in every nation. Technologies have, thus, been able to break into the minds of both militaries and law enforcement authorities by achieving a more and more privileged status during any planning, management and execution process of air, sea or land missions.

There are experts who even assure, without any doubt, that technoscience has always reinforced the military world based on a pervasive and peculiar version of a rationality that is masculine, mathematical, emotionless, and instrumentalist[1]. Hence, in order to understand late-twentieth-century and early-twenty-first-century warfare, it is important to comprehend how this rationality has busted into the dynamic of the act and art of waging war. From the standpoint of the collaborators of this report, it is necessary to address a subject like this not only with an interdisciplinary approach, but also from a prospective one.

Many examples of how technological advances have affected the development of weapons, means and methods of warfare, as well as of defence, surveillance, tracking and control –inside or outside of urban settlements– are studied in academies. Gunpowder, biometric control systems, intelligence production model of PCPAD[2], crossbows, combat aircraft, nuclear bombs, tanks, sophisticated platforms or drones –to name a few examples– are paradigmatic and genuine technological revolutions commonly studied by researchers on defence and security issues.

Notwithstanding, it is also important to recognise that we are stepping forward to an era in which the sciences of the artificial are going to be capable of drastically changing what we know as warfare. It is certainly difficult, if not impossible, to accurately predict the future; however, theorising, rethinking and, above all, generating well-grounded hypotheses and giving rise to argument-based discussions on the progress of humanity from a military and operational perspective are of course feasible.

In that sense, besides recognising the importance of the study of existing conventional and sophisticated weapons, a large part of the international community is increasingly interested in advancing understandings about the impacts of the use of artificial intelligence and robotics in emerging military and security technologies.

It is well known that various states are actively developing military systems that will utilise advanced technologies to assist, supplement, and, to some extent replace human soldiers in combat roles[3]. Thus, these types of systems, generally described as '*autonomous'*, are currently subject to great controversy[4] because of their promise of extensive operational changes in the conduct of hostilities in a near-future.

Thus, as already pointed out, the rapid evolution of new military technologies poses significant challenges and autonomous weapons systems are a perfect example of that reality. According to many experts on military and security matters[5], these technologies are bound to revolutionise the way wars are fought and how both national and international security are being guaranteed by governments specially through the design, manufacture and deployment of autonomous weapons and/or systems on the ground.

---

[1] See C. Hables, (1997). "The uses of Science" in "Postmodern war: the new politics of conflict", chapter four, London: Routledge. p. 72.

[2] Intelligence production model systems of PCPAD (planning and direction, collection, processing and exploitation, analysis and production, and dissemination).

[3] See T. McFarland, (2015). "*Factors shaping the legal implications of increasingly autonomous military systems*", in the International Review of the Red Cross, 97 (900), "the evolution of warfare, ICRC. P. 1314.

[4] See M. Sassóli, (2014). "Autonomous Weapons and International Humanitarian Law: Adventages, Open Technical Questions and Legal Issues to be Clarified", in the International Law Studies Journal from the U.S. Naval War College, 90 INT'L L. STUD. 386, p. 308.

[5] R. Geiss (2016). "Lethal Autonomous Weapons Systems: Technology, Definition, Ethics, Law & Security", published by the German Federal Foreing Office, Berlin. p. 2; SINGER, P. (2009), "Wired for war: the robotics revolution and conflict in the 21st century", penguin books, London, 512 p.p.; Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, *Report,* 39, U.N. Doc. A/HRC/23/47 (Apr. 9, 2013) (by Christof Heyns) {hereinafter Heyns}; Human Rights Watch, "Losing humanity: the case against killer robots" (2012), *available* at https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots; Noel E. Sharkey, "The Evitability of Autonomous Robot Warfare", 94 International Review of the Red Cross 787, (2012); R. Calo, M. Froomkin & I. Kerr (2016), "Robot Law", Edward Elgar Publishing, Northampton, 402 p.p.; N. Bhuta, S. Beck, R. Geiss, H. Liu & C. Kreb (2016). "Autonomous weapons systems. Law, ethics, policy". Cambridge University Press, London. 384 p.p.

In any case, as will be seen in this report, while fully autonomous systems do not yet exist, certain critical functions in military and law enforcement systems (both weapon and non-weapon-related) are already capable of operating autonomously[6]. This is a trend that will probably continue in the future, so any legal examination about them will increasingly raise several challenges too complex to solve in particular because of technical and operational issues. Many of these analysis, following current debates in the international and national arena, are raising challenges about how the sense and scope of the term *'autonomy'* must be understood and what its impact is in the interaction between human soldiers/officers–machines/systems/weapons.

Due to the interdisciplinary nature of this topic, significant debates on autonomous systems have been carried out concurrently over recent years in political, military, diplomatic, scientific and academic forums. One example can be seen through the meetings carried out since 2013 in the framework of the *Convention on Certain Conventional Weapons*[7] (CCW) at the United Nations Offices (Geneva, Switzerland). There, the international community has discussed in detail questions related to emerging technologies in the context of lethal autonomous weapons systems (LAWS).

The CCW's proceedings built mainly on informal meetings hosted in 2014, 2015 and 2016, and covered, inter alia, topics of mapping '*autonomy*', working definitions of autonomous weapons systems, and of the study of their issues related to the international humanitarian law (IHL)[8]. At the end, their mandate finished recommending to the 2016 Fifth Review Conference of the High Contracting Parties to the CCW the establishment of an open-ended Group of Governmental Experts (GGE) for an appropriate period of time which starts in 2017. This group is expected to explore and agree on possible recommendations related to emerging technologies and LAWS by considering all past, present and future proposals [9].

As a further contribution to this international discussion, the University of Barcelona, in collaboration with different Spanish and foreign institutions and universities, hosted an international workshop on the '*Sense and Scope of Autonomy in Emerging Military and Security Technologies*', which took place in February 2017[10]. This academic event paved the way to the elaboration of the present report, which contains a compendium of innovative articles written by experts on scientific, legal, diplomatic and military matters.

All papers reflect a broad spectrum of views on how '*autonomy*' can be understood in emerging military and security technologies and what their legal, technical, political and societal impacts are by building on discussions unfolded along a variety of reflective approaches. Hence, the report was divided into five articles structured as follows:

In the first paper, Maite López-Sánchez[11] thoroughly explores the relation between the rapid growth of technological advancements, which are experienced on a daily basis, and the increasing emergence of contexts in which part of the past's fiction becomes perceived today as more real. Subsequently, the writer explains why pop culture (science fiction particularly) has induced high expectations towards potential benefits of artificial intelligence, despite its long-time menacing applications, such as lethal robots (also known as killer robots) or cyberwarfare, seemed to be safely distant in time. However, that scenario seems much closer to reality now. According to the writer, this perception is shared by different stakeholders in our society, such as artificial intelligence practitioners and lawyers, who are becoming more aware of the necessity to provide a normative framework for regulating (or even banning) the development of autonomous weapon systems. Hence, the paper introduces the reader to some technical insights around the general concept of autonomy within the artificial intelligence discipline by considering its impact on the development of emerging military and security technologies. In conclusion, the writer highlights and explains some of the

---

[6] See ICRC (2016). "autonomous weapons systems implications of increasing autonomy in the critical functions of weapons", report March.

[7] *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*, as amended on 21 December 2001.

[8] Further information about meetings available at http://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument.

[9] See the recommendations for further work on LAWS which was considered by the 2016 Review Conference, as agreed on by consensus at the Meeting of Experts, available at http://www.unog.ch/80256EDD006B8954/(httpAssets)/6BB8A498B0A12A03C1257FDB00382863/$file/Recommendations_LAWS_2016_AdvancedVersion+(4+paras)+.pdf.

[10] The event presentations are available at these links: http://uboc.ub.edu/portal/Play/2f6fd071941e4f72a0d2dc32feda825b1d and http://uboc.ub.edu/portal/Play/f246bfa7809f4113ad1f68e6608354cf1d.

[11] Professor at the Faculty of Mathematics and Computer Science, and Coordinator of the Interuniversity Master in Artificial Intelligence. University of Barcelona.

most important concerns regarding the specific application of artificial intelligence technologies in the military, security and defence domain.

Following a scientific and interdisciplinary approach, Pablo Jiménez Schlegl's contribution[12] analyses software tools for the cognitive development of autonomous robots. He begins his argument by pointing out that robotic systems are evolving towards higher levels of autonomy. The article reviews the cognitive tools available nowadays for the fulfilment of abstract or long-term goals as well as for learning and modifying their behaviour. Thus, it presents a brief overview of the most salient cognitive techniques that can provide robots with a certain degree of autonomy and some kind of smart responses in front of a continuously changing world, with predictable evolutions as well as surprising contingencies.

After addressing several issues from an approach based largely on the sciences of the artificial, the report focuses on analysing '*autonomy*' from a more legal perspective. In this sense, Joshua Hughes[13] provides the reader with a magnificent study of the concept by giving a very orderly and coherent account of tools, which allow for an understanding of its meaning applied to autonomous weapon systems. Included in his article are some current methods to consider '*autonomy'* in such systems as well as a proposal of additional questions of analysis from the perspective of their role in military operations. Finally, the paper argues that many issues related to compliance with the law of armed conflict are raised by usage of weapon systems in roles which they cannot perform without human assistance.

Another significant contribution to this report is Migle Laukyte's[14] paper, which brings up a pragmatic and critical study about legal personality and autonomous weapons. The writer points out that the research and development of autonomous weapons is not going to stop: autonomous weapons are a way to maintain military advantage, something that the most powerful and developed countries want to preserve and cultivate. Thus, she urges us not to forget that questions on autonomous weapons should not cloud the bigger issue of war *per se*, where neither killing nor the war are neutral. Also, highlighting that one of the most complex challenges in addressing what autonomy means is terminological, the expert closes her exposition arguing that despite the existence of advantages in the use of autonomous weapons, the current discussion about their impact is increasingly complex because it lacks constructive approaches.

Finally, closing this compendium, Maziar Homayounnejad[15] presents in a comprehensive and detailed manner two common themes. Firstly, he maintains that '*autonomy'* is a term of art that must be narrowly and specifically applied to weapon systems if it is to be useful in a LAWS context. In that sense, he recognises in advance that arriving at a satisfactory mapping of '*autonomy*' and its impact is more than merely an academic exercise, because the way we understand and apply the term will have significant jurisdictional consequences. Afterward, the writer brings us to a working definition of LAWS, focusing on what it is about 'weapons autonomy' that may call for these systems to be delineated, and subject to certain additional requirements in both IHL and arms control. Secondly, the paper reflects on what kinds of weapon systems are likely to emerge as LAWS, and on the unique challenges posed by them. Throughout the article, he emphasises that the underlying purpose for which weapons autonomy and LAWS are defined is to delineate systems that may need to be subject to: a) deployment restrictions, b) stronger and additional precautions in attack (in IHL) and/or c) commonly agreed rules to promote strategic stability (in arms control).

Lastly, due to a special request by the editor of this report, Homayounnejad offers us a summary and a few remarks on Paul Scharre's presentation, *'The Terminator Dilemma'*[16], which was delivered during the international workshop. We are grateful for his gesture of extra collaboration with the compendium.

From a military and strategy perspective, Scharre's presentation begins by considering that the basic technology to build autonomous weapons that could select and engage targets on their own is here today. It affirms many governments still are not sure if we human beings should build this kind of technologies as it probably would be a bad idea because of legal, ethical or moral reasons. However, many '*adversary states*' and '*non-state actors'* are unlikely to be so concerned in that sense. Therefore, it would be worth asking: could autonomous weapons or systems give a decisive advantage to an enemy or

[12] Robotics Department Head in the IRII (Institute of Robotics and Industrial Informatics) of the Spanish National Research Council (CSIC) at the Polytechnic University of Catalonia, Spain

[13] PhD Student, Lancaster Law School and the Richardson Institute, Lancaster University, United Kingdom.

[14] A CONEX-Marie Curie Research Fellow of the Department of Private Law at the University Carlos III of Madrid, Spain.

[15] PhD Candidate, Dickson Poon School of Law, King's College London.

[16] Paul Scharre is a Senior Fellow and Director of the Future of Warfare Initiative at the Center for a New American Security. To follow the content of his presentation just access the link highlighted in the footnote No. 6.See P. Scharre (2017). "Ground Robotics: Preparing for Disruptive Change", presentation published by the Center for a New American Security, available at http://www.dtic.mil/ndia/2017/groundrobot/Scharre.pdf

delinquent? If so, can military and/or law enforcement authorities afford to fall behind? Many answers to questions like these also can be found in this section.

With that addendum, the report ends its initial proposal by bringing up a transdisciplinary and prospective overview about the realities and myths, advantages and benefits, dangers, uncertainties and risks involved in the use of robotics and artificial intelligence in research, development and innovation of emerging military and security technologies made by agents from both the public and private sectors.

Certainly, the debate and opinions in this direction continue to evolve. Despite some operational problems that have hindered GGE[17] meetings in 2017, it is undeniable that, as was shown in the discussions at the CCW former meetings, the international community is keen to resolve these matters. In any case, future discussions about these technologies should proceed from the clear understanding that, where these kinds of systems are at issue, some human control must be retained. Permitting the opposite could mean crossing through certain thresholds that may put humanity at risk.

Against this background, the present compendium offers a number of useful features –explanations of specialised terminology, statistical and scientific data, and bibliographic references, to name just a few– to expose readers of advanced researches in the field. It also works as a guide to elaboration on national, regional and international policies that must consider technical, legal and political aspects of the concept of autonomy in technologies, devices and systems for defence and security purposes. Hence, both domestic and global policy-makers as well as academics interested in the broad field of autonomous weaponry will find the contributions contained here to be of immense value in elucidating a debate that promises to gather pace over the next few years and beyond.

Any approach in the direction of this subject should be taken by the international community without falling into tricky arguments which tend to anthropomorphise machines. All weapons, means and methods of warfare, or of law enforcement, are simply our tools. They are not human peers, nor will they be fellow officials or soldiers, no matter how well-integrated they are in our security and defence apparatus[18].

Reflecting deeply on the sense and scope of autonomy from an open perspective must be to military and security experts as well as to lawyers, scientists or diplomats, the first step towards a strategic understanding of the potential risks, dangers, challenges and opportunities of autonomous systems in the current world order.

For the first time, substantive decisions concerning the upcoming of these *'game-changing'* technologies may well be made in the foreseeable future. And in that sense, is the ultimate contribution that this report, namely, provide analytic and productive tools to the international community to address significative issues related to emerging technological developments on military, security and defence field.

To finalise, I would like to thank personally all the writers whom have kindly contributed with this compendium as well as to professors Dr. Jordi Bonet and Dr. Antonio Madrid because of their all support for the organisation of the international workshop which held this year at the University of Barcelona.

Milton J. Meza-Rivas*

---

[17] Open-ended Group of Governmental Experts (GGE) on emerging technologies in the area of lethal autonomous weapons systems (LAWS). For more information jus check this link out: http://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument.

[18] Statement of the UN Institute for Disarmament Research at the CCW informal meeting of experts on LAWS, 12 April 2016, delivered by Kerstin Vignard, Deputy to the Director, p. 3, available at http://www.unog.ch/80256EDD006B8954/(httpAssets)/86C96CC8C7A932DCC1257F930057C0E3/$file/2016_LAWS+MX_GeneralExchange_Statements_UNIDIR.pdf.

* Researcher fellow of the departmental section of Public International Law and International Relations of the Faculty of Law at the University of Barcelona. Member of the organising committee for the international workshop on the '*Sense and Scope of Autonomy in Emerging Military and Security Technologies'* which took place in February 2017.

Some Insights on Artificial Intelligence Autonomy in Military Technologies

*By Maite Lopez-Sanchez**

Pop culture (science fiction particularly) has induced both high expectations about the potential benefits of Artificial Intelligence as well as about its potential threads. For a long time, menacing applications of Artificial Intelligence, such as lethal robots (also known as killer robots[1]) or cyberwarfare[2], seemed to be securely far in time. However, the rapid growth of technological advances we are experiencing on a daily basis, seem to be bringing prominent threads to a much closer reality. This view is shared by different stakeholders in our society, such as Artificial Intelligence practitioners or legal experts, who are noting the necessity to provide a normative framework for regulating (or even banning) the development of autonomous weapons.

In the context of the international workshop on autonomy in emerging military and security technologies hold at the Universitat de Barcelona in February 2017[3], the aim of this paper is twofold: Firstly, it aims to introduce some technical insights revolving the general concept of autonomy within the Artificial Intelligence discipline. Secondly, it aims to highlight some concerns about the specific application of Artificial Intelligence technologies in the military domain.

Artificial Intelligence

Generally speaking, Artificial Intelligence can be described as a research area aimed at designing "intelligent artificial entities or machines". However, the academic community has outlined several dichotomies revolving its research. Thus, some scholars differentiate between: intelligent "*thinking*" vs intelligent "*acting*"; "*human-like*" intelligence vs "*rational*" intelligence; or "*strong*" Artificial Intelligence vs "*weak*" Artificial Intelligence. Although other dichotomies exist (such as "symbolic" vs "sub-symbolic" or "centralized" vs "distributed"), we will briefly comment on these because they diverge on their overall objective (rather than on the means to achieve it):

- Firstly, some research on Artificial Intelligence put the stress in intelligence associated to *thinking*[4], a process that can be conducted at an abstract level (for problem solving, or logical reasoning, for example); whereas others study embodied entities that *act*[5] and are situated in a world (as it would be the case of robots, that inter*act* within our real world).
- Secondly, the research focusing on "*human-like*"[6] intelligence considers human intelligence to be the object to mimic or reproduce artificially (and advances in this field are expected to improve our knowledge about natural intelligence), whereas "*rational*" intelligence[7] pursuers just consider it as

* from the Volume Visualization and Artificial Intelligence Research Group, Mathematics and Computer Science Faculty. Universitat de Barcelona. E-mail: maite_lopez@ub.edu. Thanks to Milton Meza-Rivas for helpful comments on an earlier draft. The views contained in this paper are mine and not necessarily endorsed by the reviewer.

[1] Human Rights Watch and the Harvard Law School International Human Rights Clinic, *Losing Humanity: The Case against Killer Robots* (2012), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots (last visited July 10, 2017).

[2] R.A. Clarke, *Cyber War*, HarperCollins (2010).

[3] Video of the workshop available on-line at http://uboc.ub.edu/portal/Play/2f6fd071941e4f72a0d2dc32feda825b1d.

[4] '[Artificial Intelligence is the automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning…' in R. E. Bellman, *An Introduction to Artificial Intelligence: Can Computers Think?* Boyd & Fraser Publishing Company, San Francisco (1978).

[5] 'We define AI as the study of agents that receive percepts from the environment and perform actions' in S. Russell and P. Norvig, *Artificial Intelligence, a modern approach*. Prentice-Hall (1995).
'AI [..] is concerned with intelligent behaviour in artifacts.' in N. J. Nilsson, *Artificial Intelligence: A New Synthesis*. Morgan Kaufmann, California (1998).

[6] '[Artificial Intelligence is] The art of creating machines that perform functions that require intelligence when performed by people.' in R. Kurzweil, The Age of Intelligent Machines. MIT Press, Cambridge, Massachusetts (1990).

[7] 'A system is rational if it does the "right thing" given what it knows' in S. Russell and P. Norvig, *Artificial Intelligence, a modern approach*. Prentice-Hall (1995). p 30.

inspirational concept (as birds loosely inspire plane design) and expect machines to provide optimal solutions and to be more objective, effective and efficient than humans are (for instance, we expect a route planner application to provide the best directions to reach our destination).

- Thirdly, *strong* Artificial Intelligence (also known as True Artificial Intelligence or Human-Level Artificial Intelligence) aims at producing machines as skilful as average humans in all their facets (such as solving complex problems, running, cooking, or understanding jokes), whereas *weak* Artificial Intelligence (a.k.a, narrow Artificial Intelligence) specializes in *one* area and may well outperform best humans (such as Deep Blue, which won Garry Kasparov –a chess world champion-- in 1997 or, more recently, AlphaGo, which won against Lee Sedol --a Go world champion-- in 2016).

It is worth mentioning that these alternative dimensions are not self-excluding nor complete, and they involve many interesting issues such as consciousness[8], emotions[9], or morality[10] that are currently being studied by the research community. Other issues, such as artificial superintelligence (and the singularity[11]) correspond to a more speculative future.

Considering a historical perspective, since its creation in 1956, Artificial Intelligence has experienced a series of cycles where a hype epoch is followed by a "*winter*" where disappointment and funding cuts last for several years (or even decades) only to be broken by the renewed interest that brings next hype in the area. These cycles usually relate to one remarkable technical breakthrough in the "*narrow*" Artificial Intelligence sense that causes too high social expectations on the advance towards the "*strong*" Artificial Intelligence. Nevertheless, despite these cycles, many thousands of Artificial Intelligence applications are deeply embedded in the infrastructure of every industry[12], and thus, we should try to discern between "hypes" and real sound advance.

In this vein, Katja Grace (from Oxford University's Future of Humanity Institute) and her colleagues from AI Impacts and the Department of Political Science from Yale University, surveyed 352 active machine-learning experts to gather their guess about the number of years it would take for Artificial Intelligence to reach key milestones in human capabilities[13]. Nevertheless, before getting into the details of this study (detailed in Section 0) we should give a general intuition about machine learning and its relation to autonomy.

On intelligent algorithms, autonomy and learning

Within Computer Science, Artificial Intelligence has provided a myriad of formal specifications (mathematical problem descriptions) and algorithms so to have artificial entities to show an intelligent behaviour (see footnotes 4,5,6, and 7, for alternative Artificial Intelligence definitions) be it an smart washing machine or a highly sophisticated airport scheduling system.

Typically, we think of an algorithm as a sequence of actions or programming instructions, so that, when executed by a computer, solves a given problem (e.g., to sum up two numbers, finding the cheapest plane ticket over the web, or controlling the settings of a nuclear plant). Most algorithms are programmed by humans to solve specific problems given some inputs (e.g., considering previous examples: the numbers to add, trip destination, or maximum heat). Although algorithmic verification is an active theoretical research area, most algorithms are tested practically, checking that, when executed with different inputs, they actually provide the correct results. Thus, following the previous addition example, inputs [2,3] should return 5, whereas [2, -3] should return -1, and so on (see Figure 1). This empirical testing approach may seem enough to guarantee that the algorithm behaves as intended, but this would only be the case if we tested it with all possible inputs, and this is most often unfeasible (even in the simple case of adding numbers we may encounter some problems, such as, for example, the representation problem when adding the pi number).

---

[8] *Artificial consciousness* on-line definition at https://en.wikipedia.org/wiki/Artificial_consciousness (last visited July 11, 2017).

[9] L. Cañamero, *Emotion understanding from the perspective of autonomous robots research*. In Neural Networks, 18 (4) pp. 445-455. (2005).

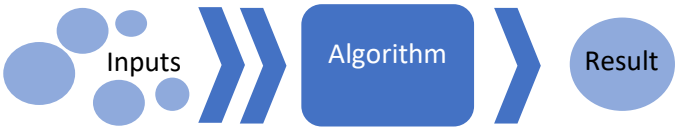
[10] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, New York (2010).

[11] R. Kurzweil, *The Singularity is Near. Viking Press (2005).*

[12] *AI winter* on-line definition at https://en.wikipedia.org/wiki/AI_winter (last visited July 6, 2017).

[13] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, O. Evans, When Will AI Exceed Human Performance? Evidence from AI Experts, arXiv:1705.08807v2 (May 2017).

The unfeasibility of extensive tests prevents us from ensuring results, but conducting enough tests can provide reasonable guarantees for most classical algorithms. Unfortunately, intelligent and autonomous algorithms usually become harder to test.



Inputs	Alg.	Result	Correct?
2, 3	Add	5	Yes
2, -3	Add	-1	Yes
-1, 1.1	Add	0.1	Yes
π, 1	Add	4.14	?

Figure 1: Algorithm empirical verification example.

Autonomy (from the greek αὐτο- *auto*- "self" and νόμος *nomos*, "law", is understood as "one who gives oneself one's own law"), when related to algorithms, is usually encapsulated in the concept of *autonomous agents*: artificial entities able to take decisions (in order to meet its design objectives) without human intervention[14]. Of course, we hardly can think of a number addition as a process that requires much decisions, and thus we would never dare to say that addition algorithm is autonomous. Nevertheless, we can think of a smart washing machine deciding the amount of water or energy to use or an intelligent wastewater treatment plant autonomously setting its functioning parameters.

Most often, autonomy is associated to robots to denote that they are unmanned, so they decide their actions without humans commanding them. Autonomy though is a concept that applies at a scale. On the one hand, industrial robots in a car factory (these devoted to attach components to the car and are fixed to the production line) are pre-programed to perform certain fixed movements and are not considered to be autonomous. On the other hand, mobile transportation robots deployed in the factory plant, do not have pre-programmed trajectories. Instead, they compute the best path to follow (having as an input a map of the plant and the information coming from their sensors). Thus, we can tell they are autonomous in deciding the path to follow, although they do it in a quite foreseeable way.

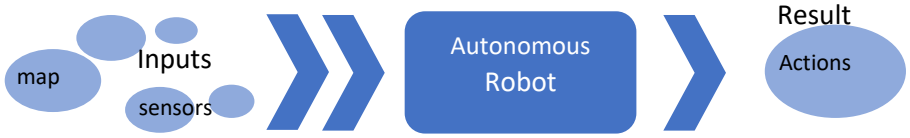


Figure 2: Illustration of an autonomous robot taking decisions on the actions to perform based on its inputs.

Clearly, autonomy and intelligence are related, since for example, planning a path in a known environment is considered to be an intelligent behaviour[15]. However, some autonomous robots are considered to be more intelligent than others if they present higher adaptation capabilities when confronting unexpected situations (such as encountering obstacles). Obviously, the diversity of contingencies that autonomous robots need to face depends on the complexity of their assigned tasks (objectives) and the scenarios where they are situated in, since, for instance, transporting an object within a factory plant is far simpler than guiding people within an amusement park.

Envisioning all possible situations that a robot may encounter usually turns out to be unfeasible. Machine Learning tries to cope with this limitation by providing computational methods to acquire new knowledge. Back in 1959, A. Samuel defined Machine Learning as giving "computers the ability to learn without being explicitly programmed". This means that the programmer specifies the sequence of instructions that will incorporate knowledge to the algorithm, rather than directly encoding this knowledge into program instructions. Consequently, the results of such algorithms depend, not only on the learning process (or code), but also on what it is actually learnt (the knowledge that is provided or the experiences

[14] (adapted from) M. J. Woolridge, *Introduction to Multiagent Systems*. John Wiley & Sons, New York (2001).
[15] P. Jiménez, Software tools for the cognitive development of autonomous robots. This publication, 18

that it acquires), so that its results are less foreseeable than those from algorithms encoded with pre-defined knowledge. Next section is devoted to providing some intuitions on how alternative Machine Learning approaches work.

Machine Learning

Learning possibly constitutes the human capability that is most intrinsically associated to intelligence, and, in fact, it may not be by chance that Machine Learning[16] is the research area within Artificial Intelligence that is getting most attention nowadays.

Generally speaking, Machine learning is devoted to algorithms that learn from data. They do not constitute standard algorithms because they do not follow a pre-defined sequence of instructions. Instead, they build a model from sample inputs (inputs that are representative of what is aimed to be learned) and use this model to make predictions or decisions. Usually, these algorithms are used when there are no standard algorithms that can solve the same problems effectively or efficiently.

There are many different machine learning algorithms. However, they all follow a typical learning-exploitation scheme. First, a learning phase is conducted so that a model is built from input training data. Afterwards, an exploitation phase comes where decisions can be autonomously taken when considering new input data. As an example, imagine the problem of discerning between cats and dogs. As Figure 3 shows, we may provide the learning algorithm with images of both animals. In the learning phase, the algorithm will extract the image features that are distinctive for both animals (that is, it will create the model based on input data), and then, in the exploitation phase, when we provide the algorithm with a new image, it will answer (i.e., autonomously decide) if it is a cat or a dog. As another example, consider a robot that faces the problem of finding its way in an unknown environment. The learning phase will consist of the robot trying out different actions in the environment (the inputs being the sensory information from the environment and an occasional valuation of the success on getting to its target) so that it builds a model of the environment and related desirable actions (a kind of map with action directions). Then, in the exploitation phase, the robot will be able to reach its target effectively.

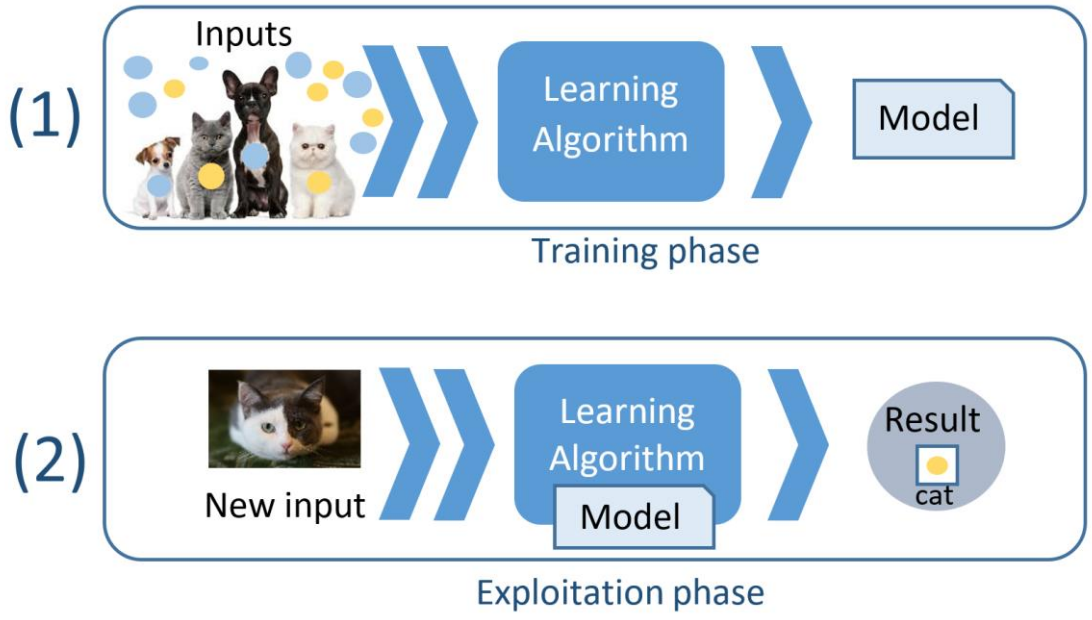


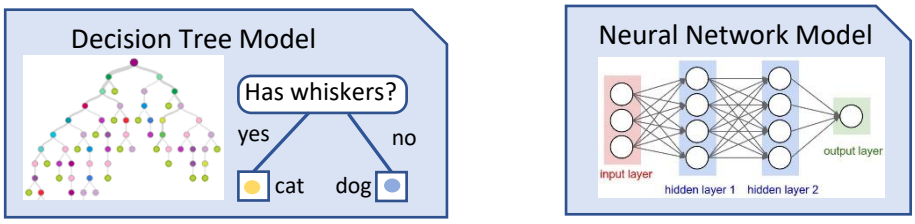
Figure 3: Illustration of the two learning phases in the cats & dogs distinguishing example.[17]

[16] *Machine learning* on-line introduction: https://en.wikipedia.org/wiki/Machine_learning (last visited July 11, 2017).
[17] Cats and dogs' images from http://more-sky.com and https://pixabay.com/ respectively.

Machine learning algorithms vary in their input data, the nature of the models they build, and the problems they are aimed to solve.

Firstly, if their input data is provided by a "*teacher*" and constitutes a representative collection of problems and their solutions, those algorithms are considered to perform *supervised learning*. Otherwise, if the input data lacks solutions, we refer to *unsupervised learning*. Considering again the cats & dogs example, the learning processes is supervised if the cat and dog images used for the training phase are provided together with its correct labelling (in fact this is the case in Figure 3, where each cat image is labelled with yellow circle and dog images are labelled with blue circles). In this case, at the exploitation phase, when a new (not learned) image is provided, the system uses the learned model to assign a cat or dog label to this new image[18]. As for unsupervised learning algorithms, they take as input the whole set of images (without associated labels). In this case, the algorithm will discover that these images can be separated in two different classes (class 1 and class 2) without any external guidance[19]. Hopefully these classes will correspond to cats and dogs, but it could instead autonomously create a model that distinguishes images of animal faces from pictures of complete animals or perform any other grouping based on, for example, animals' hair colour. In any case, in the exploitation phase, when a new image is provided, the algorithm will use the model to output one of the learned classes.

Secondly, the model they produce can be built in terms of meaningful symbols or not (technically it is said to be *symbolic* or *sub-symbolic*). The former uses explicit abstract symbols to capture the knowledge in the model whereas the later does not. *Decision Trees*[20] and *Probabilistic Graphical Models*[21] constitute some examples of symbolic models (see left hand side of Figure 4). Nowadays, *Neural Networks*[22] (see right hand side of Figure 4) are the most representative example of sub-symbolic methods, where the resulting models can be seen as black boxes: sophisticated technical machinery that, when provided with a new input, produces an output without providing any details about the reasons for doing so. Still with the cat and dog distinguishing example, symbolic approaches would consider predefined animal features, such as the presence of whiskers, to build the model. Thus, a model could include knowledge of the form "if the animal has whiskers, then it is a cat"[23]. This knowledge can be used for self-explanatory purposes, so the system could justify its classification of an animal as a cat because it has whiskers. Alternatively, when a new animal image is provided to a *sub-symbolic* model, it will return the corresponding animal classification (whether if it is a cat or a dog) without any meaningful explanation of its decision. There are some incipient efforts[24] to open this back box, but its sub-symbolic nature makes it an intrinsically hard problem to solve. Figure 5 illustrates this idea.



[18] In terms of autonomy, this means the algorithm decides if the image corresponds to a cat or to a dog. In terms of intelligence, the algorithm is smart enough to distinguish between cats and dogs.
[19] Some unsupervised learning algorithms group input data based on locations, distances, or densities.
[20] L. *Rokach and O. Maimon, Data mining with decision trees: theory and applications. World Scientific Pub Co Inc (2008).*
[21] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press. (2009).
[22] C. M. Bishop, Neural networks for pattern recognition. Oxford university press. (1995).
[23] Obviously, this assumes the model lacks knowledge about any other animal species having whiskers, since it just considers cats and dogs.
[24] M. D. Zeiler and R. Fergus, *Visualizing and Understanding Convolutional Networks in ECCV2014.* Part I, LNCS 8689, 818–833 (2014).

Figure 4: Illustration of alternative models for the cats & dogs example. Left: simple symbolic model based on Decision Trees that classifies an animal as a cat if it has whiskers or as a dog if otherwise. Right: representation of a sub-symbolic neural network model composed by 4 connected layers of neurons[25].

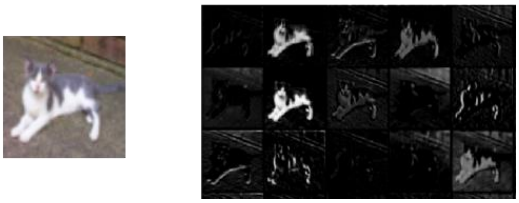


Figure 5: The image of the right shows different visualizations of the output of a given layer of a Convolutional Neural Network when it is provided, as input, with the cat image on the left[26]. Although neurons' activations eventually yield to a cat classification, the reasons for this classification are not apparent.

Finally, different machine learning approaches also differ in the problems that are aimed at solving. On the one hand, for example, *classification*[27] methods are aimed at predicting the classes (from a predefined set of classes) that correspond to new inputs. *Clustering*[28] methods are the ones that discover input groupings (or classes) whereas *regression*[29] methods do not output discrete classes but continuous numerical values. Thus, instead of predicting if an image corresponds to a cat or to a dog, if an incoming e-mail is spam, or if some bank transactions constitute financial fraud (all these being classification examples), it can be used to determine the loan that a bank will make to a customer. On the other hand, when learning to act within an environment, the problem at hand involves determining the best action to perform at each given situation. Thus, the application of a learned model results in a sequence of actions. Robots and game players do confront this kind of problems, and quite often they use a specific method called *reinforcement learning*[30]. Related to this approach, artificial intelligence optimization methods, such as *genetic algorithms*[31], are aimed at finding the best solution to a problem, be it a scheduling for an event or a design of an electrical circuit. It is important to notice that the way we can restrict the possible outcomes of all these different methods varies to a large extent. As for *classification* methods, we may not be able to foresee the class that will be assigned to a new input, since the error of the classification will depend on the representativeness of the training data. Nevertheless, we can ensure the system will always output one of the predefined classes, independently of the classification error. Thus, following the cats & dogs example, if we provide a new image, we can ensure the algorithm will (correctly or not) classify it as a cat or a dog even if it is a rare canine or feline specimen[32] or even if it belongs to another species such as, let's say, a fox. Regarding the output of methods that learn to act, these methods are typically deployed with some predefined basic actions to perform from which they do not deviate (i.e., algorithms do not invent new actions). Nevertheless, there exists a large amount of possible combinations of the sequence of actions that can result from their learning process. Thus, for example, given a fixed set of basic actions such as move forward (F), turn right (R), and turn left (L), the number of possible combinations of these actions (e.g., R-F-L, F-F-R-F-L, R-F-F-L-F-F-…) is just limited by the available resources such as time or energy. Knowing in advance the consequences of learning and executing action policies may become as hard as trying them all.

This leads us to a further discussion on testing. In fact, deciding the trade-off between training and exploitation remains as an open issue in the research community. Usually, developers decide when to switch from training phase to exploitation phase by considering some sort of threshold. But some questions arise: Do we have the time or computational resources to ensure learning? In case we had, how could we assess if the system had learned enough? These questions relate to the quality and representativeness of the

---

[25] Decision Tree image from BigML https://bigml.com/ and Neural Network image from Andrej Karpathy http://cs231n.github.io/neural-networks-1/ (last visited July 13, 2017).
[26] Images from Cameron Mence http://www.subsubroutine.com/sub-subroutine/2016/9/30/cats-and-dogs-and-convolutional-neural-networks (last visited July 12, 2017).
[27] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas. *Supervised machine learning: A review of classification techniques* in Emerging Artificial Intelligence Applications in Computer Engineering. IOS Press. 3-24 (2007).
[28] *Clustering* on-line definition https://en.wikipedia.org/wiki/Cluster_analysis (last visited July 12, 2017).
[29] *Regression* on-line definition https://en.wikipedia.org/wiki/Regression_analysis (last visited July 12, 2017).
[30] P. Jiménez, Software tools for the cognitive development of autonomous robots. This pupblication, 18
[31] *Genetic Algorithms* on-line definition https://en.wikipedia.org/wiki/Genetic_algorithm (last visited July 12, 2017).
[32] Rare individuals such as the one depicted in http://www.express.co.uk/news/nature/730827/Dog-cat-incredible-pet-viral-online-baffles-animal-lovers are confusing even for manual (human) classification.

training data. So, following the cats & dogs example, it is not clear what pictures of dogs and cats do we need to provide to the algorithm so that whenever it is confronted to a new picture, it is able to provide the solution correctly. Of course, we can compute some statistics of how many new pictures it classifies correctly from a set of a testing images, and we can even decide to stop training whenever a certain proportion of positive results have been obtained. But it will always be the case that the quality of the model (i.e., knowledge) that the algorithm has learned will depend on the quality of the data (and most of the times it is hard to assess how representative is this data with respect to the complete set of all possible images of cats and dogs). Additionally, algorithms have the risk to replicate the bias that input data may have. This bias may be intentionally introduced (to promote some options over others) or may be naturally added by the data source or the way data is gathered. In any case, it will influence the output, which will suffer from a lack of objectivity[33].

At this point it may be worth also commenting on the relationship between Machine Learning and the different dimensions of Artificial Intelligence first section introduced. The two examples provided before (a program that discerns between cat and dog images and a robot that learns how to find its way in an unknown environment) respectively illustrate cases of intelligent *thinking* and intelligent *acting*. Thus, Machine Learning is general enough to cover both approaches. Regarding the "*human-like*" vs the "*rational*" intelligence, Machine Learning mostly takes the rational approach, since it tries to provide the best solution[34]. Even those methods, such as *Neural Networks*, that may seem to be close to human brains, are in fact mathematical abstractions of brain neurons. Finally, as Machine Learning provides rather general learning algorithms, it is often prompted as the most promising approach towards *strong* Artificial Intelligence.

Deep learning[35] is a *supervised sub*-symbolic machine learning technique that has recently become a real breakthrough in areas such as computer vision (e.g. image recognition) or natural language processing (e.g. automatic translation). This is not only due to the algorithm itself but the increasing computing power combined with the huge amount of training data that internet has made available.

Although current machine learning advances could be seen as "*narrow*" intelligent behaviours, some voices claim that the composition of the myriad of services available on the cloud (be they based on machine learning or not) could constitute the practical path towards an emergent general intelligence. Other practitioners also claim that the lack of self-explanatory capabilities of unsupervised and "black-box" methods could be compensated by gaining trust from the users[36]. This is based on the idea that humans are inclined to trust those systems that provide positive results repeatedly. Nevertheless, application domains have to be necessarily taken into account, since reduced error percentages may be still be very relevant if they have fatal consequences. Recall, for example, the infamous case of an Iran air flight being shot down and resulting in 290 civilian deaths[37].

Future expectations

As aforementioned, Artificial Intelligence brings (or has the potential to bring) about many advances and threads. However, Artificial Intelligence progress, which relies on scientific breakthroughs, becomes intrinsically hard to predict and thus it is a highly speculative endeavour to try to forecast if or when this will actually happen.

With the aim of being as informed as possible, several surveys of expert opinion have been conducted[38]. From them, we highlight the one mentioned in first section from Katja Grace from Oxford

---

[33] (adapted from) Z.-H. Zhou; *Machine learning challenges and impact: an interview with Thomas Dietterich*. National Science Review. Oxford Academic nwx045. doi: 10.1093/nsr/nwx045 (2017).

[34] However, those previously mentioned biases that input data can induce in the models can also be seen as a reproduction of human biases.

[35] J. Schmidhuber, *Deep learning in neural networks: An overview*. Neural networks, 61, 85-117. (2015).

[36] J Pearson, *When AI Goes Wrong, We Won't Be Able to Ask It Why* posted this article on July 2016 https://motherboard.vice.com/en_us/article/vv7yd4/ai-deep-learning-ethics-right-to-explanation (last visited July 13, 2017).

[37] Ticonderoga-class cruiser https://en.wikipedia.org/wiki/Ticonderoga-class_cruiser shooting down Iran Air Flight 655 https://en.wikipedia.org/wiki/Iran_Air_Flight_655

[38] S. D. Baum, B. Goertzel, and T. G. Goertzel. *How long until human-level AI? results from an expert assessment*. Technological Forecasting and Social Change, 78(1):185–195, (2011);

University's Future of Humanity Institute and her colleagues from AI Impacts and the Department of Political Science from Yale University[13]. In 2016, they surveyed 352 active machine-learning experts[39] from 43 countries to gather their guess about the number of years it would take for Artificial Intelligence to reach key milestones in human capabilities. According to their answers, researchers predicted Artificial Intelligence will outperform humans in many activities in the incoming years, writing high-school essays (by 2026), driving a truck (by 2027), working in retail (by 2031), writing a bestselling book (by 2049), and working as a surgeon (by 2053). Interestingly enough, researchers expected language translation will be accomplished by 2024, which, in light of current advances, appears to be a rather pessimistic prediction. Nevertheless, we can hardly extrapolate this deviation. Furthermore, researchers report to believe there is a 50% chance of Artificial Intelligence outperforming humans in all tasks in 45 years and of automating all human jobs in 120 years. Nevertheless, authors report a large inter-subject variation, since Asian respondents expect Artificial Intelligence will outperform humans in all tasks in 30 years, whereas North Americans expect it in 74 years. Surprisingly enough, both citation count or seniority did not have an impact in this prediction.

Survey respondents saw as likely to have positive outcomes from Artificial Intelligence advances but they also saw as possible catastrophic risks. In fact, forty-eight percent of respondents expressed that research on minimizing the risks of Artificial Intelligence should be prioritized by society more than the status quo. In fact, risk may deserve particular attention when considering sensitive artificial intelligence applications such as the military ones. Next section tackles this issue.

Autonomy and learning in Military Technologies

Artificial Intelligence can be applied to virtually every field, and armed conflict is no exception. What makes it special is the ethical concerns and social alarm it raises. In fact, war has been historically highly regulated by means of international laws and agreements (such as the International Humanitarian Law) due to its highly sensitive nature.

When focusing on Artificial Intelligence, Autonomous Weapon Systems (AWS) is a term that has been coined to refer to a weapon able to make a decision/choice without human intervention[40]. Similarly, the term LAWS stands for Lethal Autonomous Weapons Systems[41]. Specifically, an Autonomous Weapon System has been formally defined as a weapon system that, once activated, is capable of selecting and engaging targets without further intervention by a human operator[42]. However, current conceptions of AWS range enormously. On one end of the spectrum, an AWS can be an automated component of an existing weapon. On the other, it may correspond to a platform that is itself capable of sensing, learning, and launching resulting attacks.

Many arguments have been made for and against autonomous weapons. A claim in favour is that replacing human soldiers by machines will reduce casualties for the owner, whereas detractors consider that it will lower the threshold for going to battle. Overall, there is some consensus on the fact that they should be regulated or even banned. In fact, as for June 2017, 3105 AI & Robotics researchers and 17701 others (including Stephen Hawking, Elon Musk and Steve Wozniak have endorsed an Open Letter[43] to ban autonomous weapons arguing that "[t]he key question for humanity today is whether to start a global Artificial Intelligence arms race or to prevent it from starting." Those endorsing the letter "believe that AI has great potential to benefit humanity in many ways, and that the goal of the field should be to do so. [But, the letter also cautions] If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, [and these weapons] will become ubiquitous and cheap [not only for] military powers [but they will be also very attractive in black market for terrorists and dictators, since] autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a

V. C. Müller and N. Bostrom. *Future progress in artificial intelligence: A survey of expert opinion*. In Fundamental issues of artificial intelligence, chap. 4, 553–570. Springer, (2016).

[39] A majority (82%) of surveyed experts worked in academia, while 21% worked in industry.

[40] D. Lewis, G. Blum, and N. Modirzadeh, War-Algorithm Accountability (August 31, 2016). Available at SSRN: https://ssrn.com/abstract=2832734.

[41] Government of Switzerland, Towards a "compliance-based" approach to LAWS [Lethal Autonomous Weapons Systems] 1 (March 30, 2016) working paper submitted at Informal meeting of experts on (LAWS) http://www.unog.ch/80256EDD006B8954/(httpAssets)/D2D66A9C427958D6C1257F8700415473/$file/2016_LAWS+MX_CountryPaper+Switzerland.pdf (last visited July 12, 2017).

[42] US Department of Defense directive 3000.09. November 21, 2012.

[43] Open letter available on-line at https://futureoflife.org/open-letter-autonomous-weapons (last visited July 12, 2017).

particular ethnic group." This letter promulgates a ban in line of that of the Biological Weapons Convention stabilised in 1972 or the Outer Space Treaty from 1967 banning space-based nuclear weapons, which also were broadly supported by chemists, biologists and physicists.

Considering legal initiatives, we may highlight "The Campaign to Stop Killer Robots" led by Jody Williams, Nobel Peace Laureate, which advocates for banning them[44]. Others claim that normative framework is needed to regulate them on international basis. Some efforts have already happened, since 2014 there have been several editions of the Convention on Conventional Weapons (CCW) Informal Expert Meeting on Lethal Autonomous Weapons Systems[45]. It is commonly agreed that, when considering Autonomous Weapons Systems in armed conflict, the law which applies is international Humanitarian Law (IHL), [46] specifically the rules relevant to targeting.[47]. in armed conflicts, IHL establishes a number of requirements for the targeting systems. Subsequent subsections are devoted to discussing on the feasibility that such requirements are met by current targeting systems.

Engaging military objectives

IHL requires targeting systems to determine if they are engaging (attacking) military objectives. Autonomy thus is related to deciding if a target actually corresponds to a military objective. Military objectives are defined as "objects which by nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage".

A prototypical example of a military objective would be an attacking tank. Recognising a tank automatically from an image would represent a sophisticated version of our previous cat & dog distinguishing example. As aforementioned, image object recognition is a machine learning research area that has greatly improved in the last years. Pinterest or Google Images[48] provide reverse image search[49] applications that allow us to experience this advance by searching for similar images from the web. These applications combine Information Retrieval methods with advanced learning techniques such as *deep learning* (convolutional neural networks) typically used in object recognition. As for today, these image methods still have unresolved issues[50], but machine learning experts are optimistic about their advances and engineers envision applications combining information from different sources (not only images) to increase the accuracy in the recognition task[38,51]. However, even if detecting a tank seems to be a task that may become feasible in the short term, this only covers the object itself. Assessing that this tank is actually attacking means to be able to analyse its circumstances, and this constitutes a far more complex problem. *Long Short Term Memory Networks*[52] are a kind of Neural Networks conceived for dealing with sequential data (such as video, which is no more than a sequence of images) that may be of help here, but it is far from being clear how can we train such a system.

---

[44] Stop killer robots' website: https://www.stopkillerrobots.org/ ;
International Human Rights Clinic (IHRC, Harvard Law School): http://hrp.law.harvard.edu/tag/campaign-to-stop-killer-robots/ (last visited July 12, 2017).

[45] http://www.unog.ch/80256EE600585943/(httpPages)/37D51189AC4FB6E1C1257F4D004CAFB2

[46] If it is not the case, a completely different law, such as International Human Rights Law (IHRL) may apply.

[47] Duke's Center on Law, Ethics and National Security (LENS) annual national security conference on *Autonomous Weapons in the Age of Hybrid War* https://law.duke.edu/lens/conference/2016/ session recordings available on-line at https://www.youtube.com/watch?v=b5mz7Y2FmU4&feature=share (last visited July 12, 2017). Prof. Michael A. Newton (28').

[48] https://www.google.com/intl/es419/insidesearch/features/images/searchbyimage.html (last visited July 12, 2017).

[49] Reverse image search allows to search the web (i.e., to formulate the search query) by providing an image instead of keywords. https://en.wikipedia.org/wiki/Reverse_image_search (last visited July 12, 2017).

[50] First ever death in an autonomous car happened on May 2017 when a Tesla car, operating in autopilot mode, hit an articulated lorry because it failed to distinguish the white truck against the brightly lit sky https://www.newscientist.com/article/2095740-tesla-driver-dies-in-first-fatal-autonomous-car-crash-in-us/. (last visited July 12, 2017).

[51] A.R. Dombe, *Biometric Target Recognition* Israel Defense. Posted on line 16/3/2017 (last visited July 12, 2017). http://www.israeldefense.co.il/en/node/28881

[52] S. Hochreiter and J. Schmidhuber. *Long short-term memory.* Neural computation, 9(8), 1735-1780. (1997).

Military objectives also include objects that the enemy intends to use. Activity recognition[53] constitutes another Artificial Intelligence research area that can be of help here (in fact, the term Activity-Based Intelligence[54] has already being coined for its application to Military Intelligence Analysis). Most approaches in this area are *symbolic* and can range from *logical reasoning* to *probabilistic reasoning*, which explicitly represent uncertainty in reasoning about possible plans or goals in formalised scenarios. Formalised scenarios are logical (*symbolic*) specifications that partially describe real scenarios (i.e., they apply simplifications to tackle the inherent complexity/intractability of real scenarios). Complete applicability in the short term to real scenarios seems to be unfeasible. Furthermore, if considering learning, it would require conducting training and testing phases that may be difficult to accomplish in many armed conflict scenarios. In fact, defining the trade-off of training and exploitation in an armed conflict scenario stands for deciding when to start engaging (attacking) a target. Necessarily, we should have the possibility of disabling destruction capabilities during a testing phase (since killing someone while testing if an autonomous lethal weapon works is immoral). Nevertheless, deciding when to fully deploy such lethal weapon in a real scenario amounts to pose the questions raised in Section 0 related to assessing if the system has learned enough and if it has been tested enough.

Similarly, assessing that an object "makes an effective contribution to military action and whose neutralization offers a definite military advantage", poses an extremely complex problem. A possible approximation would require to describe conflicting situations formally (or, at least, symbolically), so that different areas in Artificial Intelligence, such as *rule-based reasoning* methods[55] or *non-deterministic* (i.e., probabilistic[21]) predicting methods, may be combined to assess the future contributions of an object or the advantages of neutralizing it. Alternatively, one may consider that *Reinforcement Learning*[30] implicitly implies to discover the gain in performing certain actions (that is the value of the consequences). Although promising, the learning phase of this method takes so much (and so long) of trial and error that would not be feasible (and again, immoral) to apply in a battle scenario. It is worth mentioning, though, that roboticists face similar problems and they use simulations for "cheaper"[56] robot training. Nevertheless, despite the existence of visually realistic war games, we still lack of the required large variety of real world models of battle scenarios. This is especially the case if we think of the myriad of situations that may involve civilians and their properties in an armed conflict.

Finally, it is worth devoting some thoughts about the part of the sentence above that reads "in the circumstances ruling at the time". Determining if a tank is attacking or if it is surrendering can be considered to be part of these circumstances. Assessing if the tank still is a military objective or if the lethal autonomous weapon system should accept its surrender, is not only a matter of having intelligent algorithms with high discernibility capabilities. Instead, we need to consider the underlaying values that we, as humans developing such algorithms, should be able to instil in them.[57] They should be moral values aligned with those that support the normative framework that has led to regulations such as the International Humanitarian Law. Considering moral values is a relatively incipient area in Artificial Intelligence in general, and in multi-agent systems (a research area focused on the interactions among several intelligent systems) in particular. Much further research and development needs to be conducted before we can guarantee that systems will act in accordance with fundamental human values such as clemency or empathy.[58]. In fact,

---

[53] "Activity recognition aims to recognize the actions and goals of one or more agents from a series of observations on the agents' actions and the environmental conditions" extracted from https://en.wikipedia.org/wiki/Activity_recognition (last visited July 12, 2017).

[54] '[Activity-Based Intelligence] will aid in the development and understanding of patterns of life, which in turn will enable analysts to differentiate abnormal from normal activities as well as potentially defining a "new normal."' by C.P. Atwood http://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-77/jfq-77_24-33_Atwood.pdf (last visited July 12, 2017).

[55] *Rule based system* on-line definition at https://en.wikipedia.org/wiki/Rule-based_system (last visited July 12, 2017).

[56] Cheaper meaning here less cost in terms of, not only computational resources and energy, but also of more important losses and risks, such as casualties.

[57] Duke's Center on Law, Ethics and National Security (LENS) annual national security conference on *Autonomous Weapons in the Age of Hybrid War* https://law.duke.edu/lens/conference/2016/ session recordings available on-line at https://www.youtube.com/watch?v=b5mz7Y2FmU4&feature=share (last visited July 12, 2017). Prof. Michael A. Newton (55').

[58] 'My research hypothesis is that intelligent robots can behave more ethically in the battlefield than humans currently can' declarations by R.C. Arkin, a computer scientist at Georgia Tech when interviewed in Nov 2008 by C. Dean from the NYTimes: http://www.nytimes.com/2008/11/25/science/25robots.html (last visited July 12, 2017).

Moral machines is a term coined for describing the need for robots to discern between right and wrong[59]. If considering autonomous agents (not only hardware, but also software), then we can find other initiatives that are also pushing this issue. Ethicaa[60], an organisation focused on ethics and autonomous agents, constitutes an example. In general, the artificial intelligence research community is getting more concerned about moral issues, and some other terms and concepts such as *responsible AI*[61], or the Moral Turing Test[62] are flourishing.

Lawful systems and lawful behaviour

The International Humanitarian Law also requires targeting systems to be lawful and to follow lawful tactics, which entails a lawful use and lawful behaviour of autonomous weapons. Thus, it must be ensured that, once deployed, an autonomous weapon will not exhibit certain behaviours such as, for example, shooting civilians. Autonomy in this case implies the capability of discerning between combatants and non-combatants (which do not only include civilians but also different *hors de combat* soldiers such as, for example, wounded combatants or fighter pilots parachuting from their disabled aircraft). As for previous problem of military objective identification, this specific task can be really hard to achieve on its own. Needleless to say that ensuring the system will not exhibit any other possible unlawful behaviour may well be equivalent to verify a learning system (recall again previous section 0 on Machine Learning, which points out the inherent difficulties in foreseeing all possible behaviours that a system may exhibit when learning to act in an unknown environment).

Autonomy for a lethal weapon will also imply being able to adapt to changing circumstances. The most evident case would be when, after the weapon's deployment, a combatant gets injured and becomes non-combatant. However, we can think about many other situations requiring an accurate analysis about the evolution of the target context in the conduct of hostilities during an armed conflict. This leads us to two key concepts: precautions and proportionality.

On the one hand, precautions in attack relate to the required precautions that the system should take in order to minimise potential harm to civilians (or civilians' objects). Currently, attack commanders or system operators do have the obligation to exercise constant care to ensure the autonomous system does not harm civilians (nor civilian objects). This implies having humans on the loop.

On the other hand, under the International Humanitarian Law, the proportionality principle holds that the autonomous weapon cannot engage with a target if the expected harm that it will cause on civilians (or civilian objects or a combination thereof) is excessive compared to the anticipated military advantage. If conducted intentionally, a disproportionate attack may constitute a war crime.[63]

Being able to act without violating these key principles requires autonomous weapons to show an intelligence that appears to be closer to that pursued by *strong* artificial intelligence (see section 0) rather than to current narrow intelligent approaches. As we have discussed in section 0, human level artificial intelligence may not be that far in the future, but to put it in a timeframe would be highly speculative. Unfortunately, this does not prevent some countries, such as South Korea or Israel, from deploying armed robot border guards[64]. However, it is important to keep in mind that physical weapons may be most visible, but software intelligent systems have the potential of being far more ubiquitous. Next section makes a brief remark on this.

Autonomy in the cyberspace.

As previously mentioned, physical systems (e.g. killer robots) have a limitation in their deployment in real scenarios mainly due to the difficulties of perceiving and being aware of the exact circumstances of

---

[59] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right From Wrong*, Oxford University Press (2010).
[60] Ethics and autonomous agents http://www.ethicaa.org/ (last visited July 12, 2017).
[61] Responsible Intelligent Systems website: https://responsibleintelligentsystems.sites.uu.nl/ (last visited July 12, 2017).
[62] J. Parthemore and B. Whitby, *What Makes Any Agent a Moral Agent*? International Journal of Machine Consciousness. 05, 105. https://doi.org/10.1142/S1793843013500017 (2013)
[63] Extracted from definition of proportionality in attacks (under IHL) fom the Geneva's Accademy's Weapons Law Encyclopedia http://www.weaponslaw.org/glossary/proportionality-in-attacks-ihl (last visited July 12, 2017).
[64] D. Cavanaugh, *Robot guns guard the borders of some countries, and more might follow their lead* posted to Offiziere.ch on April 12, 2016. https://www.offiziere.ch/?p=27012 (last visited July 12, 2017).

the hostilities. However, cyberspace, the environment in which communication over computer networks occurs, is based on structured networked infrastructures and data. Thus, cyberspace constitutes the natural context for intelligent algorithms as well as for any other computer science technologies such as *distributed architectures* or *security*, since they base their functioning in data and/or communication.

Distributed systems are especially relevant technologies that enable the processing of massive volumes of data. However, distributed systems do also have the disadvantage of being detrimental for the assignment of responsibilities. In fact, accountability represents a rather slippery issue for most learning processes, since the final results depend on may factors (e.g., programming, acquired experiences, or deployment circumstances). However, matters become worse when these results do not come from the computation of a single entity but from many[65].

Security (or cybersecurity) is meant to protect from unintended or unauthorized access to technological equipment and services as well as to protect from appropriation, change or destruction of data. In brief, it is the area that studies how to prevent and handle cyberattacks (i.e., attacks within cyberspace). Cyberattacks can range from installing spyware[66] on a personal computer to attempts to destroy the infrastructure of entire nations. In the context of armed conflict, cyberwarfare involves both offensive and defensive operations pertaining to the threat of cyberattacks, espionage and sabotage.

Cyberattacks and espionage constitute threats that have been installed in our society for some time. State and cyberterrorists develop cyber weapons that are employed against specific targets and meet objectives which would otherwise require espionage or the use of force. Some voices have been raised against government developing software for unlocking electronic devices claiming that it would put people and countries at a greater risk since it will weaken the overall security and privacy if it fell into the wrong hands.

As for many other application areas, Artificial Intelligence is presented as the future of Cybersecurity. Unfortunately, Artificial Intelligence can not only be used for protection but also in undesirable cyberattacks[67]. Indeed, machine learning methods for extracting information have the potential to surpass by far current cyberattacks. Likewise image object recognition, there has been a huge advance in the research area of Natural Language Processing. Thus, for example, Google's neural machine translation system[68], which uses a refined version of a Long Short Term Memory network[52], provides state of the art translations. But before this actually happens, some voices are already urging governments to invest more in counter-cyberterrorism and revisit their policies in stockpiling and securing cyber weapons, such as those which were leaked from the CIA and are being used as ransomware in medical facilities around the world[69].

Conclusions

This paper tackles some of the issues that arise when considering Artificial Intelligence applied to military technologies. Firstly, it introduces general notions of Artificial Intelligence in order to illustrate that intelligent algorithms, although fitting in the general algorithm definition of a sequence of programming instructions, are particularly hard to verify (i.e., to assess its correct performance). This is specially the case for algorithms with learning capabilities, whose results not only depend on the absence of errors in the actual programming, but also on the characteristics of the training settings (e.g., how representative they are or if they are biased) and deployment circumstances.

Very closely related to intelligence appears the concept of autonomy, rooted on decision making. Thus, the paper defines autonomous systems as those able to make decisions (in order to meet its design

---

[65] This is not only the case for virtual cyberspace, since, for example, there are deployments of multiple physical robots that coordinate and constitute a swarm https://en.wikipedia.org/wiki/Swarm_robotics.

[66] Spyware: software that aims to gather information about a person or organization without their knowledge, that may send such information to another entity without the consumer's consent, or that asserts control over a device without the consumer's knowledge.

[67] R. V. Yampolskiy *AI Is the Future of Cybersecurity, for Better and for Worse*. Posted to Harvard Business Review on May 2017. https://hbr.org/2017/05/ai-is-the-future-of-cybersecurity-for-better-and-for-worse (last visited July 12, 2017).

[68] W.Wu, M. Schuster, Z. Chen, Q.v. Le, M. Norouzi et al. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation* arXiv.org > cs > arXiv:1609.08144 available at https://arxiv.org/abs/1609.08144

[69] http://houseofbots.com/news-detail/758-1-what-ceos-should-know-about-cybersecurity Posted by N. Kishor on July 4th 2017.

objectives) without human intervention. In this manner, objectives become key for understanding the behaviour of such systems. It is worth noticing, though, that objectives can be noble (desirable) as well as mean (undesirable) or any combination thereof (e.g., desirable goals that turn out to be not so good), and this is the basic rationale behind the advances and threads that Artificial Intelligence brings about. It is also important to have in mind that objectives may be changed (or even hacked if the system is not secured enough) at different stages of systems' design, training or deployment.

When considering armed conflict settings, Artificial Intelligence applications become highly sensitive, since decision making may involve casualties. This necessarily requires legal and moral analysis and action. Voices from the social, science, technology, ethical, and legal communities have been raised to put a stop to the artificial intelligence military race. Other voices advocate for a proper regulation aligned with the International Humanitarian Law. This paper analyses some aspects that illustrate that the current technology is still far from being able to comply with legal requirements.

To conclude, a final statement should be done to highlight the necessity of a breakthrough in both minimizing the inherent risks of intelligent algorithms as well as in including moral values in decision-making processes.

Software tools for the cognitive development of autonomous robots

Pablo Jiménez *

Abstract

Robotic systems are evolving towards higher degrees of autonomy. This paper reviews the cognitive tools available nowadays for the fulfilment of abstract or long-term goals as well as for learning and modifying their behaviour.

Introduction

Machines are powered artefacts intended at performing a given action. They can be viewed as more or less sophisticated tools, designed for executing specific work. Among all the machines developed by human mind, robots deserve a special place. The standard definition (ISO 8373) of an industrial robot stresses the following key aspects of its unique nature:

- Automatic control: robots deploy their activity without intervention of a human.
- Multifunctionality: robots are not constrained to perform a single, specialized action, but the same robot arm can execute as different tasks as manipulation, painting, welding, or inspection, with a simple (possibly automated) tool change.
- Reprogrammability: the trajectories and tasks developed by the robot can be easily modified by software, i.e., rewriting and/or executing a different program, without the need of readjusting the hardware.
- Continuous workspace: the robot can position itself in any point within its reach and follow arbitrary trajectories.

These features distinguish industrial robots from conventional machines. However, the degree of autonomy of such robots is heavily restricted by their environment: they are only able to deal with incidences within a quite constrained and structured world. They are certainly able to adjust their trajectories to the actual position and orientation of arriving workpieces, or to react to failures like the absence of a piece supposed to be there or to observed defects. They can even respond to certain non-trivial sensory input like computer vision, e.g. for classification or failure detection. But to cope with the contingencies of the unstructured world outside controlled environments, some more steps have to be taken. First, sensory input becomes now an unavoidable must. The world is continuously changing, and agents like robots need to have an updated picture of the world's state in order to be able perform meaningful actions. Basic home robots like vacuum cleaners rely on very simple bug-like sensor systems based on infrared range measuring and contact detection. This suffices for the duties they are endowed with. The environment is now not as structured as in the case of industrial robots, but it has still well-defined features: a home vacuum-cleaner robot is not expected to be deployed on the streets, and a lawn-mover robot will rapidly be stuck in the savannah (if not attacked by a cheetah!). Moreover, the work they have to perform is quite simple and specific.

More advanced interaction with the surroundings like object manipulation or cooperation with humans requires a quite more sophisticated perception. Computer vision techniques are to be applied for image processing, object (or people) recognition, and scene understanding (i.e., annotating its elements and establishing their –mostly spatial– relationships). In many cases, the robot will need to learn what it is watching at. Learning in perception is in first instance a classification or categorization issue: when not

* Institut de Robòtica i Informàtica Industrial (CSIC - UPC), Llorens i Artigas 4-6, E-08028 Barcelona, Spain. e-mail: pjimenez@iri.upc.edu Work finished April 3, 2017. This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under project RobInstruct (TIN2014-58178-R).

predefined by the human through a set of salient features, classes have to be constructed from a number of labelled examples

whose attributes have to be generalized so that in a later recognition phase the specific object perceived is correctly fit within its corresponding class. The classified object is automatically bestowed with the characteristic attributes of its class. Nowadays, with tools like the Convolutional Neural Networks (CNNs) computer vision is quite reliable in classification, even with regard to visual hindrances like poor illumination or partial occlusions. However, artificial systems are still far from human performance with respect to interpreting a scene, especially if a number of common sense or cultural cues are involved.

Perception alone does not suffice, the robot must also be able to decide which action to perform next and to execute this action. Decision making in Robotics is obviously intended to be an autonomous process. The scope of the required cognitive load ranges from purely reactive to highly deliberative systems. These different approaches are briefly reviewed in Section “Decision making in Robotics”. They provide the framework within which the cognitive processes that are needed to grant autonomy to the robotic agent take place. The term cognitive process has to be interpreted as referring to those algorithmic procedures that mimic higher mental processes in humans, without pretending the machine to own a mind.

Short-term decisions taken by a robot are not quite difficult to trace (i.e., to follow the process or to understand how they have been taken) or even to predict, in general. More involved are long-term decisions, as they rely on a planning process that may include a high number of planning operators (PO) that represent the individual actions, and a much higher number of possible combinations of such POs. These POs may be either provided as a fixed repertoire by the human designer or programmer, or they can be learned by the robotic agent directly from its perception channels. This said, it is clear that the two relevant cognition processes involved in decision making are learning and planning. Despite learning takes place earlier in the information processing than planning (we will see that this is not always true), we will examine planning first (Section “Planning”) in order to get a better idea of what has actually to be learned, and then some learning paradigms will be overviewed in Section “Learning”. These two sections constitute the core of this presentation, and besides providing a brief description of the corresponding processes, we will emphasize to which extent their outcomes are determined by human design or intervention. A summary and some reflections on the latter are finally provided in the concluding remarks of the paper.

Decision making in Robotics

The behaviour of robots is managed by the control system, like in any automatic machine. In the case of robots, this control is heavily software-based. The computer programs that decide in any moment how the robot will respond to specific stimuli and how it will perform its duties have been designed and encoded by human programmers and users. The different control schemata can be summarized as in[1]:

- Reactive control ("don’t think, react")
- Deliberative control ("think, then act")
- Hybrid control ("think and act independently in parallel")
- Behaviour-based control ("think the way you act").

Purely reactive controls are also the most predictable ones -up to failures-, as any response to foreseen inputs has been previously programmed by the human, and unforeseen inputs are simply ignored. The main advantage of this control architecture is its swift response, its main draw-back its lack of flexibility. Deliberative systems, on the contrary, thanks to their powerful cognitive tools -which are the subject of the rest of this contribution-, display a far more elaborated adaptability to changing conditions of the environment. Nonetheless, their response may come too late for a system which is embedded in the real world with all its threats and contingencies, as such deliberative processes are highly time- and resource-consuming. This justifies the hybrid control architecture, which provides immediate response to urgent issues (as long as they are predictable) thanks to the reactive component, whereas they are also

[1] M. Matarić, ‘Learning in behavior-based multi-robot systems: Policies, models, and other agents’ (2001) 2(1) Cognitive Systems Research 81

capable of long-term adaptation to the world changes (sometimes such systems are also said as being both goal and event driven).

As for behaviour-based control, it will not be tackled in this paper, but it deserves some words before proceeding to the cognitive processes. Behaviours can be thought of as small programs encoding input-output responses, operating at different levels of abstraction (without making any hierarchical structure explicit) and highly interconnected. The famous behaviour-based subsumption architecture of Rodney Brooks is the paradigmatic example of this control concept, postulating that complex behaviours can emerge from this interconnectivity and activity of simple behaviours. Predictability goes lost as the complexity of the system grows. The idea of emergent behaviour appears also in swarm robotics, each robotic unit being quite simple but the whole swarm being highly interconnected. Thanks to this dense communication and feedback, simple rules can make emerge complex behaviours.

As just said, we will concentrate on the cognitive processes of deliberative systems. Although they are presented independently from one another, it should be clear that reasoning constitutes in some sense the theoretical background of the other functions, and planning and learning are intimately related, as the trajectories or the symbolic actions used by the planner may be the out-put of the learner. Moreover, planning may be part of the learning process (at symbolic level): the planner uses the rules as learned so far to determine whether they are enough and correct to al-low the accomplishment of the task. If not, the necessary modifications in the symbolic representation of actions have to be undertaken (rule refinement). Thus, planning does not only provide the required instruction sequence for task fulfilment (to be translated into robot commands) but also plays an active role in the learning process.

Knowledge representation and reasoning

Reasoning occupies the highest level in an intelligent robot's control architecture. Reasoning operates on a certain kind of knowledge representation (KR), which in Robotics may belong to one of three types:

- Logic-based KR. Statements are either true or false, and knowledge about the world's state is assumed to be complete.
- Probabilistic formulations. Statements are either true or false, but their particular truth value may be unknown.
- Fuzzy logic formalism. Instead of true or false, a statement may be true up to some degree. Stated differently: while in conventional logic the truth value belongs to the set {0, 1}, in fuzzy logic it can be any value within the interval [0, 1].

Next, the main features of these KR and associated reasoning types are presented

Logic

Logic, and more specifically First Order Predicate Logic (FOPL) is the archetypal KR in Artificial Intelligence, where reasoning is based on the powerful deductive mechanism (some reasoning systems may use inductive, abductive, or other types of logic inference mechanisms as well). Under simplifying assumptions, logic formulations and its derivate action planning mechanisms (see Section "Deterministic action planning") have been used in Robotics with some success. However, Robotics intrinsically comes with two hard problems for logic formalism. The first one is that robots are physical agents embedded in a changing world, whose actions are responsible of some of these changes. In logical terms, they have changed the truth value of some facts about the world, while others remain unchanged. Determining what changes and what remains may not be a trivial issue (this is known as the frame problem, see also again Section "Deterministic action planning"). The second point is that sensor information may conflict with previous beliefs, i.e. state a contradictory truth value about a known fact. Mechanisms may be foreseen for providing plausible explanations and resolve such conflicts, but again, this is not trivial to resolve. Such problems render a logical formulation of robots acting in the world as undecidable in general. Practical

solutions consist in considering holding periods for formulas (the notion of situation), or in sacrificing completeness[2].

Description logics (DL) have been consolidating their suitability for structuring semantic knowledge about the world. By inheriting some of the features of classical KR formalisms like semantic networks and frame systems, they not only provide semantic structure and consistency to a particular domain, but also some form of inference, and thus they can be seen as forming "a certain family of decidable subsets of FOPL".[3] DLs entail two main components:

- the upper ontology or terminological knowledge (TBox), that is, the set of concepts of a particular domain and the relationships between these concepts, in particular equality and subconcept or superclass-subclass relation, the latter enabling property inheritance by creating a hierarchical taxonomy. Concepts are unary predicates, and concept conjunction, disjunction and negation may be used as combining operators. Roles are binary predicates that allow to express a semantic relation between two concepts.
- individual objects or assertional knowledge (ABox), for grounding concepts and roles.

DLs come with some reasoning basics like consistency of the concept definition, subsumption and disjointness of concepts, consistency of the ABox with respect to the TBox, concept and role instances, all of which are decidable in a DL.[4] DLs are explained in more detail in a number of works.[5] DLs are at the base of web ontology languages, like OWL, devised at developing the semantic net. As for Robotics, serious efforts have been made to derive an ontology that allows sharing and exchanging knowledge between robots about objects, tasks and environments, like KNOWROB, based on DL, which uses OWL and exploits its hierarchical structure of classes that allows inheritance.[6] This has been extended with meta-information about the data to be exchanged, algorithms that were used for creating data and requirements that are needed for interpreting it, aiming at robots-to-robots sharing of knowledge across a robotic World Wide Web in the ROBOEARTH project.[7] In a top-ranked international professional association like the IEEE Robotics and Automation Society, a Working Group on "Ontology for Robotics and Automation" is active since 2012 and have developed already a standard on this subject.[8]

Probabilistic formulations

As embedded creatures in the world, robots rely on sensors as the main source of information about their surroundings. However, too often such information is noisy or incomplete, leading to lack of information. But almost as frequently, this does not mean an absolute ignorance about a given world state, but some kind of knowledge about the chances of the different alternatives is available in general. This can be formalized quantitatively with probabilities associated to each option, and dealt with using tools like Bayes' rule. This is an inference mechanism for computing the probability of certain event, given the priors and dependent relevant probabilities. The practical implementation of this mechanism, that allows to know the probability of a certain cause being behind the observed effect, while avoiding the huge computational cost of a naive use of this mechanism, is known as Bayes networks (BNs), and for systems evolving with time, dynamic Bayesian networks (DBNs).

[2] J. Hertzberg and R. Chatila, 'AI Reasoning Methods for Robotics' in Springer Handbook of Robotics (2008)
[3] Hertzberg and Chatila, n.2
[4] Hertzberg and Chatila, n.2
[5] See for an introduction F. Baader, I. Horrocks, and U. Sattler, 'Description Logics' in Steffen Staab and Rudi Studer (eds), Handbook on Ontologies (Springer Berlin Heidelberg 2004)
[6] M Tenorth and M Beetz, 'KNOWROB knowledge processing for autonomous personal robots' (October 2009)
[7] M Tenorth and others, 'Representation and Exchange of Knowledge About Actions, Objects, and Environments in the RoboEarth Framework' (2013) 10(3) IEEE Transactions on Automation Science and Engineering 643
[8] For more information, please see http://standards.ieee.org/develop/wg/ORA.html

Fuzzy logic

Fuzzy logic allows to reason about qualitative and approximate statements. It can be seen as a generalization of propositional logic with continuous truth values along the interval [0, 1], and reformulating logical junctions to operate with such numerical values (negation as the complementary to 1, disjunction as the maximum, conjunction as the minimum, etc.). Fuzzy logic knowledge bases, containing sets of if-then rules that relate fuzzy values of some variables to the fuzzy values of others, can be used for inference by forward chaining. Once the fuzzy value of a given variable is computed, it can be defuzzyfied by assigning a scalar value like the central point of the corresponding interval of possible values of this variable (if such a numerical value is necessary for the application at hand).

Planning

Medium- or long-term goals require some kind of planning. Here the absolute timescale is not so relevant as the evolution of the world's state: long-term would refer, with this precision in mind, a time span in which many changes take place. This is generally associated to the fact that the robot has to concatenate a sequence of actions to achieve such a goal. Each such action modifies the world's state, either just by changing the robot's configuration, or by altering some aspects of the surroundings.

In Robotics two main types of planning have to be distinguished: motion and task planning.

These two types occupy different levels as for degree of abstraction: task planning occurs at a formalistic, symbolic level, whereas motion planning takes place in a geometric mock-up of the real world. In fact, motion planning could be considered to connect task planning to the real execution of the action commands, as long as such actions involve the displacement of the robot (or of a part of it). Action specifications are generally qualitative, whereas motion guidelines are quantitative, and despite having to consider additional control issues, their translation into executable motion commands is rather straightforward.

In the next sections the main traits and variants of the two types of planning are overviewed, in order to provide a rough idea of how they work and to which extent human intervention conditions the outcome of the planner.

Motion planning

The motion planning problem is formalized as given an initial (or start) and a final (or goal) configuration of the robot, and a description of the free-space (or, in a complementary fashion, of the present obstacles), to determine a collision-free motion from the start to the goal. The notion of configuration space (C-space) is quite useful: the dimensions of such space correspond to the degrees of freedom of the robot, and therefore in such a space the robot becomes a point, and the solution path is unidimensional. The counterpart is that the planning space has now as many dimensions as the number of degrees of freedom of the robot, and the layout of such a space is not quite intuitive. C-obstacles are all the configurations resulting in a collision of the robot with surrounding objects.

Construction of the planning space

In academic toy examples, a geometrical description of the environment where planning takes place is already assumed to be provided. Other sources of existing environment descriptions are architectural plans, city maps, roadmaps, geographical maps, etc. The problem is that such maps often do not provide the level of granularity or detail required by the robot, and more importantly, they almost surely do not reflect the real layout as for the presence of obstacles. Furniture is displayed at a fixed position in an architectural plan, but chairs can be displaced and occupy unsuspected locations. A country map may not show the fences crossing a path, not to speak from all the mobile obstacles encountered in urban or rural environments, etc. Therefore, if such information is used, it has to be complemented with online observations of the environment by the robot. Aerial photographs or videos taken by a drone may provide

such updated information to ground robots. Alternatively, mobile robots equipped with online cameras and/or range sensors can con-struct their own maps, following the techniques generically known as Simultaneous Localization and Mapping (SLAM). Robust algorithms exist nowadays for solving this task.

C-obstacles are difficult to construct explicitly, besides very simple cases. The randomized motion planning algorithms explained below, however, avoid this step and resort to efficient collision-detection algorithms only when needed.

Sensor-based motion planning

It could be the case that no geometric description of the environment is available at all. Moreover, the robot may be equipped with just quite simple onboard contact or range sensors. Even with such limitations it is possible to derive some strategies that allow the robot to follow a path leading to the goal. They are known as bug-algorithms, because such robots really act as simple animals with limited perception. The robot is assumed to have some notion, at any location, of where the goal is, and proceeds towards it in a straightforward fashion and surrounding obstacles found on its way.

Classical motion planning

These methods operate on full descriptions of the C-space, i.e. a geometrical model of the space is available. They include Potential functions, Roadmap methods, Exact and Approximate cell decompositions. These methods really work well only for simple low-dimensional settings. Practical methods, working efficiently for real robots, rely on probabilistic approaches, and have to sacrifice completeness for efficiency. They are shown next.

Randomized motion planning

This family of methods bases its success on sampling the C-space and computing possible robot-obstacle collisions only at the sampled configurations, as well as at some points of the segments joining them. These methods are said to be probabilistic complete: if a solution path exists, the algorithm will eventually find it (the probability is higher the more samples are used). The two main families of methods are the Multi-query planners where the constructed roadmap can be used for different queries (i.e., finding the path between different start-goal point pairs), the paradigmatic approach being the probabilistic roadmap method (PRM), whereas the Single-query planners consider exclusively a specific start-goal pair, constructing on the fly a tree-structure for this planning query, with the Rapidly-Exploring Random Trees (RRT) as the basic algorithm of this family.

Beyond basic motion planning

Particular problems go beyond the basic formulation of motion planning, and specific methods have been devised for each of them.

- Differential constraints. Planning with constraints on velocity and acceleration is known as kynodinamic planning (in particular, planning both a path and velocities along it for a robot arm is termed trajectory planning), and if such differential constraints cannot be integrated into derivative-free constraints -like in vehicles with limited turning radius-, we have a nonholonomic planning problem.
- Multiple robots. In a scenario composed of multiple robots, collision-free paths have to be found that allow each one of the robots reach its individual goal. Decoupled approaches, like prioritized planning or xed-path coordination, are preferred to costly centralized solutions.
- Moving obstacles. It is assumed that the motion of the obstacles is known in advance. In such cases, an additional temporal dimension could be added to the configuration space, with the constraint that only forward paths along this dimension are allowed. Planning in such a space is

computationally hard. Alternatively, the problem may be decoupled into a path planning and a motion timing part.

- Manipulation planning. Here transit and transfer modes have to be considered, the first being standard motion planning problems of the robot towards a part, the second being the robot carrying the part. The achievement of stable grasps has also to be included in the planning.
- Assembly planning. Planning the ways the different parts of an assembly can be brought together, respecting the precedence constraints between the parts (some part must be mounted before others
- Planning with sensing uncertainty. This type of planning copes with limited knowledge about the configuration space. Sensor information is employed to plan in an information space instead, with information feedback about the current state.

Task planning

Task or action planning consists in symbolic planning in terms of statements about the world and the robot. Actions modify the current world state where they take place, and planning aims at sequencing or concatenating actions such that starting at an initial state, a goal state is achieved.

This concept is transversal to scheduling, which means resource allocation (time, energy consumption, etc.) to a set of actions, so that specified deadlines are met while respecting resource limitations.[9] Planning techniques with time constraints allow to cope with the two problems simultaneously, and not in cascade as traditional approaches. However, here we will concentrate on the planning problem alone.

Deterministic action planning

First planners like STRIPS (developed by Richard Fikes and Nils Nilsson in 1971 at the Stanford Research Institute for computing simple plans for the mobile robot Shakey) were based on a propositional logic formulation of the world. States are described by sets of conditions (propositional variables), so that the initial state entails the set of conditions that are true at the beginning (all the others are assumed to be false), and the goal state the set of conditions that have to be true plus the set of conditions that must be false. The planning operators (PO) represent actions, and are represented by a quadruple that includes two sets for the preconditions (conditions that must be true and conditions that must be false in order to execute the actions), and the postconditions or effects of the action (again a set of true and another of false conditions). Planning consists then in determining a sequence of POs that change the world successively from the initial to the goal state. This planning language is deterministic in that after execution of each action, effects hold completely.

This has inspired what nowadays are known as Planning Domain Description Languages (PDDL). Algorithms based on PDDL make the planning problem tractable by resorting to simplifying assumptions, that may include:

- finiteness (the domain has only finitely many objects, actions, and states)
- information completeness (the planner has all relevant information at planning time)
- determinism (actions have deterministic effects)
- instantaneousness (actions have no relevant duration)
- idleness (the environment does not change during planning).[10]

These languages extend the propositional nature of STRIPS by upgrading to predicate logic formulations, that is, allowing the existence of non-grounded variables, besides the constants, in the describing conditions. Thus, the preconditions and effects of the POs representing actions include also free variables, and planning requires not only finding a sequence of POs but also grounding consistently their

[9] Hertzberg and Chatila, n.2
[10] Hertzberg and Chatila, n.2

variables, i.e. determining valid constant values for these variables. For this reason, a PO is now also known as action schema, it is a structure where different groundings are possible.

The PDDL is also a standard to which different specific languages adhere, where some additional features may exist like argument typing, equality handling, conditional action effects, and some restricted form of FOPL statements.[11] Temporal planning, that is, allowing actions to specify durations (thus overcoming the instantaneousness assumption) is a distinctive feature of the extension PDDL2.[12]

Planning not necessarily produces a total order or linear plan, but also partial order or nonlinear plans are a possible outcome. In such plans, a set of actions and an ordering relation between them are determined, whereas unordered actions may be executed in any sequence (in some formalisms even in parallel). Classical formulations of this partial-order plan generation start with the empty plan, which contains just the initial and the goal conditions, and iteratively introduce new actions by checking if the generated conditions respect the partial ordering relations. This strategy would lead straightforwardly towards the solution if all the actions were independent, but quite frequently it appears that the effects of a newly introduced action threaten the partial ordering in the plan attained so far. A way out of such conflicting and time-consuming interactions is to resort to subplans as planning macros, thus obtaining a hierarchical structure for planning, This is the idea behind hierarchical task networks (HTNs), were the plan is incomplete as long as there exist unexpanded (i.e., to the lowest level in the hierarchy) subplans.

Newer deterministic planners which have earned considerable success like GRAPHPLAN rely on the expressive power of planning graphs and the strength of logical inference by planning as satisfiability. As a drawback, the latter introduces the well-known frame problem, i.e., how to express changes without having to state explicitly all what remains unchanged. Alternative logic formulations like deductive logic or temporal logic have also originated quite efficient planners.

Probabilistic action planning

Modern planning approaches cope with the fact that sensors often provide incomplete information, which means that planning has to be performed under uncertainty. The standard formulation of this kind of problems is the Markovian decision processes (MDPs). To the conventional sets of states S and actions A, a new feature is added, namely that action models include conditional probability distributions for the corresponding state transitions. Typically, this means to specify for each possible effect of an action a certain probability of occurrence. The aim is to obtain a policy, that is a function that maps states into actions, and such a policy may be derived from value iteration (VI) or policy iteration (PI) algorithms. As explained later in the context of reinforcement learning, such methods aim at maximizing the overall utility of the plan (where the utility of an individual action would be the negative cost associated to this action).

Like in the case of deterministic planners with PDDL, also standards have been provided in international planning competitions for probabilistic planners: PPDDL (with the first P standing for Probabilistic) and more recently RDDL (inspired in the transition models of DBNs).

One step further is to relax the complete observability assumption of the world state. This gives raise to the Partial Observability Markov Decision Processes (POMDPs), that add to the MDP formulation an observation model: a finite set O of possible observations that the robot can perform, as well as the conditional probabilities of making a specific observation o in state s. Reformulating planning in the belief space, i.e. probability distributions over the

state space corresponding to the robot's belief of being in such state after executing action a or observing o, the same search techniques as in MDPs can be applied, namely VI or PI. As belief space is exponentially larger than state space, only quite simple POMDPs can actually be tackled in this way, although also some approximations make it more tractable.[13]

[11] Hertzberg and Chatila, n.2
[12] Hertzberg and Chatila, n.2
[13] Hertzberg and Chatila, n.2

Finally we have to stress again that generally the robots have a fixed repertoire of actions, and if no combination of such actions produces a satisfactory fulfilment of the goal, the robot ends up concluding that no plan exists. The way out is providing the robot with the capability of expanding this repertoire. This means learning new actions.

Learning

Outside of the controlled and structured environments where most robots dwelled up to now, robots have to face a world they know nothing about. Their human programmers may provide them a declarative description of some of the world's traits. Such a formal description is obviously incomplete, inaccurate, simplistic and hardly useful for mere survival. In other words, robots have to carry out their activity in a world that is

- partially known, (i.e., incomplete knowledge about the world)
- partially observable (not even relevant observations can be taken for granted), and
- dynamic (i.e., changing).

As for the latter, one should add that changes stem either from ambient phenomena or from actions performed by other agents.[14] Some of such changes can be anticipated with reasonable assumptions on expected behaviors: ambient changes may follow physical laws or established rules, whereas agent actions may adjust to the knowledge about its goals and motivations. Such knowledge may be previously encoded in the robot's knowledge base, or it must be acquired, i.e. learned. Non-coded as well as unpredictable knowledge render learning as an unavoidable requisite for autonomous robots to be deployed in unstructured environments. To this end, machine learning techniques apply.

A classical but quite informative classification of learning strategies distinguishes between supervised and unsupervised methods. In supervised learning, there is a teacher providing feedback to the system about its learning performance, by providing the correct answer after execution of the learned action or task. This can also be expressed in terms of formulating the goal of learning in terms of computing the function f that relates a given input X with an output Y , that is, Y= f (X): in supervised learning, the corresponding Y to certain X is provided by the teacher, that can supervise how in successive iterations function f is approximated increasingly well. It is also the teacher who decides when the learning system has achieved an acceptable level of performance and learning terminates. The two big families of supervised learning methods are classification methods (here, each output Y is a class or category, and function f has to correctly assign each individual X to its class) and regression algorithms (both variables are numbers and f may be an analytical function like for example in linear regression). Supervised learning can of course also take place at a symbolic level, like inductive logic programming, which aims at synthesizing a minimal logical program that provides the correct true or false values to the corresponding input variables. Popular classification methods used in Robotics include:

- Support Vector Machines (SVM), which try to maximize the gap separating the data belonging to different classes (the base procedure is linear, but the kernel trick allows for non-linear separating functions as well).
- Statistical methods like Bayesian learning (an application of the Bayes rule in order to find the probability of a certain class to be the one to which the input data belong, knowing the priors of class distribution in advance and having determined the likelihoods of the data for each class in the training phase), and its variants maximum likelihood and expectation maximization.
- Neural networks (NN), which consist in combining basic computational units, called neurons, linked by weighted connections. Each such neuron computes the weighted sum of its inputs and fires if this sum is larger than a given threshold. In the learning phase, the weights associated to each neuron input are updated until the network performs a satisfactory classification. NN allow for online learning, but they provide no insight into the classifier.

[14] including both natural episodes as well as typical –possibly regular– events in human environments, like the alternation in traffic lights

The supervised learning paradigm par excellence in Robotics is Learning from Demonstration, which is described more in detail in the homonymous Section.

Unsupervised machine learning, on the contrary, provides no information about Y to the learning system. There is no teacher and the system has to determine the implicit structure underlying the input data X. That is, it tries to model such distribution or structure, and performance can be expressed with regards to how well new data adjust to the found structure. Unsupervised learning families include clustering methods (inputs are grouped in clusters by some proximity or partitioning criterion, k-means clustering being a popular such algorithm) and association rule learning (i.e., to discover rules describing large portions of input data). In Section "Reinforcement learning" we take a closer look at this widely used and typical unsupervised learning scheme in Robotics.

In the case of a robot, due to its embodiment, embedded in physical surroundings, learning heavily relies on visual perception. Perception is the input stream from which descriptions about the current state of the world can be extracted, which in turn allows to couple sensed changes in the environment to particular actions performed by the robot. Thus, before examining learning paradigms, a brief overview on perception is provided. Unless otherwise stated, we will always refer to visual perception.

Perception

In the context of learning, visual perception is doubtless the most powerful input channel for a robotic system to obtain a description of the world. It can also be the most efficient one, because of the immediate encompassment of a whole scene, as long as the involved visual processes avoid becoming a computational bottleneck. Computer vision (CV for short) is about processing static images or a continuous video stream, and this includes, in broad terms, the following steps:

- Image acquisition (with digital mono or stereo cameras, maybe with enhanced features like range measuring, or signal measuring beyond the visual spectrum),
- Preprocessing (this includes several basic image enhancing processes)
- Segmentation (division of an image into regions, that may correspond to different objects or object parts),
- Recognition (bringing image regions in correspondence with object models or labels).

Applications of CV that are relevant for Robotics include object recognition (aka classification, that is, assigning a specific view of an object to its corresponding class, which is prespecified or has been learned), identification (of an individual object, face, fingerprint, iris, or the like), or detection (of an object, a defect, a person, etc. within an image). Specific instances of recognition like facial expression recognition or gesture recognition (including its dynamic version along a video sequence) are of particular interest in human-robot interaction. They are frequently used in the learning from demonstration context, as shown below. As for this application, another perception channel reveals as being quite useful, namely measuring the forces exerted on the arm (particularly in kinesthetic learning). To this end, either force/torque sensors mounted on the wrist are used, or force measurements at the robot's joints. Turning back to CV, the concurrence of different recognized objects/people or their spatial relationships may lead to quite basic instances of scene understanding, which however are still far from the richness of human scene understanding, due to the lack of knowledge about social and cultural cues.

State of the art tools are Convolutional Neural Networks (CNNs), which are artificial neural networks inspired in the visual cortex of animals and which model visual perception by humans (and by animals in general). They perform very well in image recognition: they are close to humans in object classification and detection (as long as images are not altered with filters, as in current popular smartphone applications), and even slightly better in fine grained classification. The labelling of the individual images appearing in the databases used for training these CNNs have been done by humans. Stated differently, the basis of classification, the implicit criteria for categorization have been established by humans. Nonetheless, projects exist nowadays to perform such categorization automatically, from the text accompanying the images in the world wide web, like captions of the photographs appearing in the news.

Learning to act

In what follows we examine the two most extended learning techniques in Robotics, which happen to be quite genuine representatives of unsupervised and supervised learning. We will not go very deep into the technical detail, and instead try to provide a general idea of how they work, with special emphasis on the role played by the human programmer.

Reinforcement learning

Reinforcement learning (RL) refers to a set of algorithms devised to obtain an optimal or near-optimal policy (action selection based on the current state), without intervention of a teacher (i.e., they belong to the unsupervised learning category). The setting is conceived as a Markov Decision Process, the current state and action selected determine a probability distribution on future states, that is, the effect of applying an action depends only on the current state where the action is applied, regardless to the previous history. Full observability of each state is assumed, although partial observability formulations do also exist. The only feedback provided to the learner comes from the environment, where the robot's actions take place. There are many RL techniques, but the common features include the following:

- set of environment and robot states;
- a set of actions A that can be executed by the robot;
- policies of transitioning from states to actions;
- rules that determine the scalar immediate reward of a transition;
- rules that describe what the agent observes.

The reward typically comes with the observation of the last transition undergone, and expresses the degree to which the resulting state (or the action leading to this state) is desirable. The reward is provided by state observations, but it is up to the designer of the RL algorithm (or the user implementing it) to decide which environment (or robot) features are used for computing the reward. Rewards express immediate satisfaction degrees, but what really guides the learning process are the value functions (aka utility functions) of the transitions (or of the states), that correspond to long-term degrees of desirability. Values are computed from the rewards of the estimated optimal course of actions leading to the final goal of the learning process. A certain state (or the transition leading to it) may have a high reward but a poor value, and vice-versa. While rewards are directly taken from state observations, values must be estimated and reestimated again and again from the sequences of observations a robot makes over its entire lifetime (i.e., from the different action courses that lead to these sequences of state observations). Some RL methods use a discount factor associated to future rewards, which allows to tune the relative influence of immediate versus long-term desirability. It should also be noted that most RL methods are stochastic approximations of exact Dynamic Programming: instead of sweeping over the whole state space, sampling of states according to the underlying probabilistic model is performed.

Value estimation can thus be seen as central to RL. Nonetheless, evolutionary optimization methods (like genetic algorithms or simulated annealing), which search the policy space directly, could be used instead. These methods do not allow to interact directly with the environment while learning, whereas value function estimation RL does, but they can be used to contrast their results with those obtained with RL. This online use of RL raises another question, namely the exploration-exploitation trade-off: exploring new, potentially more rewarding states vs. exploiting current knowledge. A typical strategy to deal with this issue (among others) is the □-greedy method, where the action currently believed to be optimal is chosen with probability 1 − □, and another random action is chosen with probability □.

Future projections of the system's behaviour may either be model-based or model-free. The model mimics the system's behaviour, it allows for simulations of possible courses of actions. An example of model-based algorithm is Adaptive Real-time Dynamic Programming. Model-free algorithms, on the other hand, do not require any knowledge about the consequences of the individual actions. Q-learning is a characteristic example of model-free RL algorithm.

Learning from demonstration

Learning from demonstration means basically that the robot is taught by performing the task to be learned "in front of" it (i.e., it is a supervised learning methodology) The teacher (generally a human demonstrator) executes several instances of the task in a way that the robot's perception system is able to follow their performance. If learning is conceived as taking place in a search space (the space of all the possible solutions to determining the correct trajectory or the correct sequence of actions to attain the goal), then learning from demonstration (LfD) allows to drastically reduce this space, either by focusing or restricting learning to a close neighborhood of the solution, or by pruning away the parts corresponding to wrong solutions (by counterexamples). LfD, aka imitation learning, is also the way of programming robots in a natural and intuitive way. The human-robot interaction (HRI) tools used to this end will be examined below, but it is pertinent to mention now one of these tools, namely kinesthetic guidance. This refers to physically guiding a robot arm along the desired trajectory by pulling and pushing it at the end-effector (or other parts of the arm). And it is pertinent to mention it here because one of the very first industrial robot programming methods consisted precisely in guiding the robot (directly, with the help of a teach pendant -a kind of wired remote-, or by driving a lightweight mock-up) along the desired trajectory, which was registered for later reproduction in the execution phase. What distinguishes LfD from these early programming ways is that not an exact reproduction of the taught trajectory is sought, but a generalization over several such demonstrations, which are executed in slightly varying conditions. As a result of the learning process, such a generalization aims at adjusting to the current conditions during execution.

Skill transfer in LfD means to answer the following questions

- What to imitate?
- How to imitate?
- Who to imitate?
- When to imitate?

The who and when questions haven't received much attention in research, as in the usual setting there is just one teacher, and the instants where the demonstration begins and ends are well-established. What to imitate refers to determine which are the relevant parts of the demonstration that need to be learned. This is achieved by the repeated demonstrations in the learning phase: only the relevant parts of the task are expected to be maintained along the demonstrations (thus, certain variability is desirable). Furthermore, and this is an HRI issue, social cues may be used for focusing the attention on the important parts of the task: gazing or pointing at region and time-intervals of interest, using verbal statements, etc. This means of course that the robot has to be previously endowed with the capacity of interpreting such social cues. Furthermore, it has to be established whether the robot is intended to reproduce the articular motions of the teacher, to follow the trajectory of the hand, or if only the final position attained is relevant. The first variant, pursued e.g. in gesture learning, may be impossible due to large differences in the embodiment of teacher and robot (see the correspondence problem below). In any case, also a metric of imitation performance has to be defined, associated to the different alternatives the imitation may be conceived, namely as achieving the same final relative position, the same absolute position, or the same relative displacement as in the demonstration.

How to imitate addresses the so-called correspondence problem, which relates to the different embodiment of teacher and learner, therefore exhibiting a different kinematic structure (such differences may be just a question of scaling, or of different proportions, or even of different number, disposition or type of joints).

In the research community, LfD is usually distinguished as occurring at trajectory level or at task (or symbolic) level, and in the previous paragraphs we have been switching indistinctively between both modalities. At trajectory level, the robot learns to perform basic sensory-motor skills, and for this reason it can also be envisioned as learning control policies. It is a generalization of movements, aiming at obtaining

a generic representation of motion which allows to encode very different types of signals/gestures.[15] The what-to-imitate problem translates into which variables have to be chosen to encode a movement. Alternatives addressed by researchers include encoding at joint level, at task level, or in torque space. The type of motion to be encoded may be cyclic (i.e., repeated under slightly varying conditions), discrete, or combining both types. Prior to the actual learning, dimensionality reduction techniques may help to reduce its computational load. The original recorded signals are projected onto a latent space with fewer dimensions while preserving the maximum amount of information. A typical such technique is Principal Component Analysis (PCA), where the main direction to be projected upon is the one where the data exhibit the highest variance, the next one is the following highest variance direction, with the constraint of being orthogonal to the first one, and so forth. The encoding may be based on statistical analysis (like Gaussian Mixture Models, combined with Gaussian Mixture Regression for the predictions (GMM/GMR), Hidden Markov Models (HMMs), or other), or based on dynamic systems, like Dynamic Motion Primitives (DMPs), Locally weighted regression (LWR) or even recurrent Neural Networks.[16]. Vision and proprioceptive sensors are the most common input sources, although force/torque sensors have also been considered in recent years,[17] especially for the cues they provide in collaborative tasks.[18] As for the latter, robots involved in collaborative tasks, the assignment and switching of the roles of master (teacher) and follower within the same task have also been studied based on interaction forces,[19] or switching between reactive and proactive behaviours by anticipating human motions.[20]

As for task level learning, the aim is to formulate the task in terms of predefined atomic actions (or small standardized sequences of atomic actions) represented symbolically, for example as rules in STRIPS-like formalisms. Such rules, if appropriately learned, can be concatenated afterwards by using a planner to reproduce the sequence of actions that fulfils the task under slightly different start and goal conditions. Such sequences of actions can also be encoded and reproduced using classical machine learning algorithms like HMMs, or graph-based hierarchical encodings.[21] Symbolic learning requires the sensory input to be processed and segmented into meaningful world transitions corresponding to the aforementioned actions. Sequencing entails learning some kind of precedence constraints between actions. Some actions have to strictly precede others, and this is discovered after a sufficient number of demonstrations whose variability uncovers such strict precedence, distinguishing them from actions that only circumstantially happen to occur in a given temporal order. While task or symbolic learning allows to learn interactively high-level skills, these methods have the disadvantage of relying on a large amount of prior knowledge (e.g. about the basic atomic actions) in order that the demonstrated sequences can be segmented consistently.

In LfD, the more demonstrations provided to the robot, the better should the task at hand be learned by the robotic system. However, the more demonstrations have to be shown to the robot, the more annoying and tedious becomes teaching the task to the robotic unit. Ideally a few demonstrations should suffice, as long as they are different enough as to highlight the really significant traits of the task, as said above. But, with the exception of very simple cases, it is difficult in general to design such a set of significant demonstrations. An interesting way out is to resort to some kind of incremental learning. Rough versions of the task are learned with as few demonstrations as possible, and the robot starts executing them right away. Performance may be poor at the beginning, but monitoring of the robot by the programmer or the user allows to detect where improvements are required. The learned task is progressively refined by providing new demonstrations. The errors observed in early performances allow the teacher to identify where the new demonstrations have to provide new insights about the task. Verbal and non-verbal cues can be used by the teacher to guide the attention of the robot system towards the

[15] A. Billard et al., 'Robot Programming by Demonstration' in Springer Handbook of Robotics (2008).
[16] M. Ito et al., 'Dynamic and interactive generation of object handling behaviors by a small humanoid robot us-ing a dynamic neural network model' (2006) 19(3) Neural Networks 323 i
[17] L. Rozo, P. Jiménez, and C. Torras, 'A Robot Learning from Demonstration Framework to Perform Force-based Manipulation Tasks' (2013) 6(1) Intell. Serv. Robot. 33.
[18] L Rozo et al., 'Learning Physical Collaborative Robot Behaviors From Human Demonstrations' (2016) 32(3) IEEE Transactions on Robotics 513.
[19] Y Li and others, 'Role adaptation of human and robot in collaborative tasks' (May 2015).
[20] W Sheng, A Thobbi, and Y Gu, 'An Integrated Framework for Human-Robot Collaborative Manipulation' (2015) 45(10) IEEE Transactions on Cybernetics 2030.
[21] A. Billard et al., n.15

parts of the task that need to be improved. This guided incremental learning is often called scaffolding or moulding of the robots' knowledge about the task, it can also be seen as a variant of coaching.

Driving the attention towards some parts of the task or specific locations of the setting is a typical application of HRI. Social cues have been investigated and applied to this end. They encode in some sense priors of the statistical learning methods, speeding up learning. Non-verbal cues include pointing and gazing, whereas verbal instructions require some form of natural language processing. Even the prosody of spoken instructions has been studied in the search of such social cues. As said above, CV techniques are needed for interpreting the gestures associated to such attention-driving cues.

Last but not least, another way of avoiding a huge number of demonstrations is transferring the refinement of action learning to an unsupervised method. This has the advantage of reducing drastically the search space of the latter, thanks to the demonstrated tasks, while avoiding to further resort on the teacher for obtaining a more accurate performance. Typically, such unsupervised task refinement learning consists in some form of RL. In the perspective is somehow complementary: a (relational) RL approach is enhanced with occasional requests to the teacher, who performs demonstrations oriented at pruning the search space significantly. The own system provides suggestions on which aspects should be covered by the demonstration. The idea behind this approach is that the time of the human teacher is much more valuable than the robot's time and thus requests should be kept at a minimum. 22

Finally, it should be stressed that some LfD schemes resort to biological analogues, being the most characteristic those models that try to mimic the functioning of Mirror Neuron Systems, which are responsible of imitation in animals.

Conclusions

In this contribution we have presented a brief overview on the most salient cognitive techniques that can provide robots with a certain degree of autonomy, and some kind of smart response in front of a continuously changing world, with predictable evolutions but also surprising contingencies. In first place we have perception, the input to the control system and thus to decision making. Perception depends on the identification parameters provided by the human designers/programmers, and modern classification algorithms like the powerful CNNs operate on labelled data (or contextual information). In other words, the semantic content has been previously given by humans. This statement can also be extended to methods learning from data extracted from internet, with the potential risks attached to such an uncontrollable source.

In simple reactive controls the response is always perfectly defined, and can be reliably predicted, up to errors or failures. The problem becomes more involved when higher cognitive functionality is invoked. Motion planners compute paths of the robots (maybe trajectories, if the different velocities along the path are also considered) and even such an apparently innocuous duty can have social or ethical implications if, for example, such a path appears traversing a sensible area. The geometrical basis of such planners excludes any moral responsibility of the robotic agent, as it corresponds to the human programmer to exclude such areas from free-space, or to foresee their possible existence if the system builds its maps autonomously, and in this case, the means for identifying such areas should be provided.

As for task planning, we have seen that deterministic planners rely on logical linking of actions. Of course this can become arbitrarily complex, but at least traceability of plans is granted. Undesired side-effects from some actions can be traced back for uncovering the conditions that produce them, which in turn may have consequences on the design of the actions (more than on the process of planning per se). Probabilistic task planners, on the other hand, are more difficult to trace back-if not impossible- due to the randomness of some decisions:[23] the probability of occurrence of each effect is naturally known, but the actual effect that finally takes place is clearly unknown until it happens. Such probabilities may provide a quantization of the liability of the human designer to each outcome, a kind of calculated risk, but the

[22] D. Martínez, G. Alenyà, and C. Torras, 'Relational reinforcement learning with guided demonstrations' [2016] Artificial Intelligence

[23] unless, of course, a register of all the history, i.e., of the whole sequence of states and actions, is kept, and not just the final result

chaining of several actions may introduce rapidly a high degree of complexity, as for the combinatorics of effects. Moreover, it is a common practice to include a number of spurious effects within a generic "noise" effect with certain probability.

We have seen the two paradigmatic learning strategies. Supervised learning, and in particular, learning from demonstration clearly assigns the whole responsibility of what is learned to the teacher, the one who provides the demonstrations. Liability may be limited by the learning performance of the system, and the errors that may arise during the learning process.[24] In unsupervised learning the issue is a little bit more tricky, although what is considered a reward to guide the learning development is of course a decision of the designer. Even the fact that it is long-term value function and not the immediate reward, as we have seen, what determines action selection, the computation of the value function still bases on what the designer has considered to be the state variables to promote. Model-based learning even allows for a certain kind of predictions.

In sum, cognitive processes in robotic systems have been designed and implemented by humans. Deterministic processes can be predicted, as for their evolution and final result, or traced back. In the case of probabilistic processes, also certain predictions on their behaviour can be made, with associated probabilities of occurrence, although combinatorics and insufficient computing power may pose some limits to the designer's or programmer's anticipation capabilities. This can be partially alleviated by performing a worst-case analysis, which reduces the options to consider. In any case, it is up to the human programmers/users to decide whether complexity and uncertainty may shelter questionable decisions of the robotic system. As for today, there is nothing like an autonomous will of a robotic system, and the morality of its actions is the morality of its human designers, programmers or users.

[24] See also. M. Lopez-Sanchez, 'Some Insights on Artificial Intelligence Autonomy in Military Technologies' for the difficulties of assigning liability to AI systems in the military context, this publication, 5.

What is autonomy in weapon systems, and how do we analyse it? – An international law perspective.

By Joshua Hughes[*]

1. Introduction

This paper discusses the meaning of autonomy in weapon systems, and how it is analysed in international law. It firstly considers what autonomy means, and how it has been studied using commonly used lenses of analysis (the 'levels of autonomy', and the 'loop' paradigms). The paper introduces some additional analytical questions which elucidate more understanding of specific issues which are most relevant from an international law perspective. The present paper is focussed on the 'how' of analysis, rather than specific discussion of issues raised.

In adding to the analysis made possible through the commonly used paradigms, this paper argues: 1) that the legal issues which are relevant to autonomous weapon system (AWS) deployments depend upon the role which the system is intended to be used for; 2) that overcoming these legal issues and complying with legal requirements depends upon the capability the system possesses; 3) as the number and/or complexity of the roles an AWS is delegated increase, so do legal issues and the required capability to comply with legal requirements.

This paper carries out a number of analyses of both current weapon systems, and hypothetical weapon systems which may appear in the future. Three scenarios are examined, two which consider current scenarios where weapons with some autonomy have already been deployed, and one hypothetical scenario which may be possible in the future.

It contends that the unlawful use of a weapon system with autonomy can occur when legal requirements are not met by the system in the role it is used for. Where systems are not capable to the required level, humans can fill these capability gaps to ensure legal compliance. Consequently, this paper argues that potential unlawfulness of using such a system is not caused by the autonomy in a weapon system, but by inadequate technological capabilities which fail to meet the required legal standards. For example, a weapon system may struggle to differentiate between civilians and combatants, thereby not complying with the law of armed conflict (LoAC, also known as international humanitarian law, and the laws of war). However, if a human were to accurately approve lawful targets, the LoAC requirement to distinguish civilians from combatants would be complied with.

2. The meaning of autonomy

Autonomy has multiple meanings in the modern world. The origin of the word autonomy comes from 17th century Greece.[1] It began as *autonomous*, which originally meant 'self-law', or something that has its' own laws.[2] Now, however, this definition only applies in relation to self-governing states. A state is subject to its own laws, and international law which it consents to (in addition to *jus cogens* norms).[3] An alternative modern meaning of an autonomous entity is something that has autonomy. Meaning that they have '*the*

[*] PhD candidate, Lancaster Law School, Lancaster University and Co-ordinator of the Richardson Institute Internship Programme, Lancaster University

[1] Collins English Dictionary Online, 'Definition of autonomy' (2017) <https://www.collinsdictionary.com/dictionary/english/autonomy> accessed 11 January 2017

[2] Catherine Soanes and Angus Stevenson (eds), *Oxford dictionary of English* (2nd edn, Oxford University Press 2005), see 'autonomy'

[3] D. Shelton, 'International law and 'relative normativity'' in Malcolm D. Evans (ed), *International law* (3rd edn, Oxford University Press 2010), 146

freedom to determine one's own actions, and behaviours.'[4] They are therefore independent, and not subject to outsider control.

If something is truly autonomous, it has total freedom to determine its' own actions, without the influence of any other entity. Are the AWS expected to be developed in the near future actually autonomous, according to this dictionary definition? The answer, is no.

AWS, otherwise known as 'killer robots',[5] are: '*A weapon system that, once activated, can select and engage targets without further intervention by a human operator.*'[6] This comes from the US Department of Defense, and is also used by Christof Heyns, former UN Special Rapporteur on extrajudicial, summary or arbitrary executions.[7] The statement '*further intervention by a human operator*' implies that there is human involvement prior to the AWS being activated, this could be a commander or operator giving an AWS orders and tasks to perform. As AWS are unlikely to have the freedom to deploy themselves and choose to launch their own operations, they are subject to the instructions of their operators. Thus, they do not have total freedom to determine their own actions and therefore are not truly autonomous in the dictionary sense.

Furthermore, AWS will be controlled by their programming, and thus subject to constraints built into it, and not totally free to determine their own behaviours. Their programming would also be subject to LoAC, and international human rights law, as any deployment of AWS would require their usage to be compliant with these bodies of law. Thus, AWS are not *autonomous* in the dictionary sense, as they are not free to choose their own actions. They will be subject to their programming, which would contain legal requirements written as algorithms, and to orders given by a human operator.

Later in this paper, we will discuss AWS and LoAC issues. International human rights law will not be discussed here, in order to keep the paper focussed on one area of investigation. LoAC applies during an armed conflict. The Tadić decision defined an armed conflict as existing: '*whenever there is a resort to armed force between States or protracted armed violence between governmental authorities and organized armed groups or between such groups within a State.'*[8]

3. The meaning of autonomy in robotics

As we have seen, AWS would not be *autonomous* within the dictionary definitions. However, autonomy is understood differently within the field of robotics. It is more related to automatic functions than to concepts of individual freedom. Sharkey offers this explanation: an *autonomous robot* is one that can operate in open and unstructured environments, and does not need to follow a strict path of action. [9] For example, if two robots were instructed to travel from point X to point Y, an *automatic* robot could only follow a pre-determined route between X and Y, and if there were any objects blocking this route, it would be unable to complete the task, as no deviations from the route have been programmed. However, an *autonomous* robot, by virtue of its ability to function in open and unstructured environments could not only determine its' own route between X and Y, but also sense the presence of objects blocking its path, and move around them.[10]

For Crootof, Sharkeys' definition of *autonomous* means that AWS already exist.[11] Present capabilities such as defensive counter-missile weapon systems do operate in the open and unstructured environment of the real world, rather than the robotics laboratory, and so can be seen as autonomous. For

---

[4] Collins English Dictionary Online, 'Definition of autonomy' (2017) <https://www.collinsdictionary.com/dictionary/english/autonomy> accessed 11 January 2017
[5] Human Rights Watch, 'Losing humanity' (Human Rights Watch 2012), 1
[6] United States Department of Defense, 'Directive 3000.09' (2012), 13-14
[7] United States Department of Defense, n.6 above(2012), 13-14; Christof Heyns, 'Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions to the Human Rights Council. A/HRC/23/47' (UNHRC 2013)
[8] *Prosecutor v Tadić (Interlocutory Appeal) Case No. (IT-94-1-A) Appeals Chamber* [1995] ICTY, 70
[9] N. Sharkey, 'Automating warfare: Lessons learned from the drones' (2011) 21(2) Journal of Law, Information & Science 140–154, 141
[10] N. Sharkey, n.9 above, 141
[11] R. Crootof, 'The Killer Robots Are Here: Legality and Policy Implications' (2015) 36 Cardozo Law Review 1837–1915, 1863-1865

example, the Phalanx Close-in Weapon System which has been installed on military bases, and naval ships, can, when instructed to do so, recognise, locate, and destroy incoming ordnance, such as rockets, artillery, and missiles before these dangers can reach the intended target. The Phalanx does this all without human intervention.[12]

However, Sharkey's definition of autonomous does not emphasise the same factors as alternative definitions within robotics. For others, they are explicit that both the absence of human involvement and the absence of pre-planned functions are required.[13] Thus, the Phalanx system would not be *autonomous*, as being programmed to solely destroy incoming ordnance, it can only perform the pre-planned function of recognising ordnance and destroying it.

Here, we can see that whilst all systems with some freedom to make decisions based upon their programming in open and unstructured environments have some autonomy (the Phalanx 'chooses' to recognise a missile as something to neutralise). Some of these systems are restricted to only performing functions which have been pre-programmed by human beings. Potentially, future AWS may be able to recognise a target and engage them, without being explicitly ordered to carry out an operation against that target. For example, a future-AWS in the form of an unmanned aerial vehicle (UAV or 'drone') with high-level sensors that recognise almost all physical entities, near human cognition, and an ability to reliably function with wide autonomy could be ordered to seek out its own targets: to find and neutralise anybody, or anything, which it recognises as the enemy. This is different from the Phalanx, where specific targets are inputted, and it is given a specific role to destroy them.

Present day targeting systems with autonomy detect objects using sensors and compare data about those objects to its on-board memory of targets it should neutralise.[14] In the future, a far more complex AWS with unrestricted movement, and the capability to recognise many more types of targets would also be doing the same, comparing detected objects to pre-programmed targets. However, this hypothetical future-AWS may have much wider autonomy than current systems, they could distinguish any enemy, such as troops, suicide bombers, or tanks from the entirety of its environment, rather than being programmed to recognise and eliminate one single type of target.

Herein lies the major legal issue with AWS when deployed against humans, in order to recognise enemy human targets they would need to compare data about potential targets and the environment from their sensors to compare to target signatures in their memory. However, this requires reducing the difference between civilian and combatant down to data. The difference may be so small where terrorists or militants are fighting in civilian clothes, that an AWS cannot make the distinction, thereby putting civilians at risk. Furthermore, the requirement to distinguish civilians from combatants or other forms of legitimate target feeds into other requirements such as reducing civilian harms and balancing that harm with military advantage – all of these issues currently require human judgement to perform, which AWS may never reach.[15] Such highly complex decision-making would require highly-advanced systems in order to achieve legal requirements, and therefore avoid potentially unlawful uses of AWS. Critics of AWS argue that such decision-making will always require innate human judgement, which machines would never be able to replicate no matter how advanced.[16]

Both current systems, and our hypothetical future-AWS would be controlled by a computer system that could be described as 'intelligent', as both operations include calculations and tasks that would normally require human cognition. We could, therefore, call them systems with artificial intelligence. However, they would be 'narrow' artificial intelligence (AI) systems, also known as 'weak AI'. This means that the AI can only perform a single task, or group of tasks, related to a single or a small number of objectives, in this case controlling a weapon system. A system that could perform all human functions would be a 'general' AI, and

---

[12] Raytheon, 'Phalanx Close-In Weapon System' (28 June 2016) <http://www.raytheon.co.uk/capabilities/products/phalanx/> accessed 6 January 2017

[13] R. Hooper, 'Robotics glossary' (*Learn about robots*) <http://www.learnaboutrobots.com/glossary.htm> accessed 11 January 2017; RobotWorx, 'Robot Glossary of Terms' <https://www.robots.com/glossary> accessed 11 January 2017

[14] B. Handy (ed.), 'Royal Air Force: Aircraft and Weapons' (RAF 2007), 87

[15] Human Rights Watch, n.5 above, 30-36; see also M. Sassoli, 'Can autonomous weapon systems respect the principles of distinction, proportionality and precaution?' in ICRC, 'Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects' (ICRC 2014), 41-43

[16] Human Rights Watch, n.5 above,32-34

a system functioning beyond human capability would be a 'super' AI.[17] This paper will not consider any possible dystopian, or utopian, futures as a result of 'super' AI's,[18] but will focus upon AWS controlled by a narrow AI.

Although both controlled by a narrow AI, operations carried out by the future-AWS creates many more legal issues than those created by the Phalanx (which have already been overcome, due to their previous deployment).[19] However, if the future-AWS were to, hypothetically, be deployed in a role similar to the Phalanx, perhaps defending a small military outpost from incoming artillery whilst it was under siege, would there still be greater legal issues? Unless it were to begin targeting people, which would be a different role, there would not be. This paper argues that the level of autonomy is dependent upon the operational role, with the number of associated legal issues increasing with the level of autonomy that is required. Before making that argument, let us first look at the rules governing the conduct of AWS in armed conflicts, and common paradigms for analysing autonomy in weapon systems.

4. Basic principles of the law of armed conflict

LoAC has three main principles in relation to targeting: distinction; proportionality; precautions in attack (this is a limited explanation of LoAC to enable the present discussion. It is not a complete account).[20]

Firstly, distinction requires that parties to a conflict will distinguish between civilians and combatants, and between civilian and military objects.[21] This becomes very difficult when civilians become involved in fighting by directly participating in hostilities through forming/joining an organised armed group, or attacking parties to the conflict individually.[22] Furthermore, combatants who are *hors de combat* (out of the fight) by becoming prisoners of war, surrender, or incapacitation through wounds or sickness cannot be targeted.[23] In addition, attacks which treat multiple targets as a single object would be indiscriminate and prohibited.[24] For example, if an enemy were to take five strategic points from which to defend a village, targeting the whole village, rather than those five points would be indiscriminate.

A key point related to distinction are the rules around doubt as to combatant status. Where an individual cannot be identified as an enemy combatant or irregular fighter, they are to be considered civilians.[25] Thus, should a future-AWS be deployed with the ability to target humans, they would not be able to target any individual unless able to identify them as having a combatant status other than civilian, i.e. an enemy combatant, or irregular fighter. Therefore, an AWS could not target its weapons at an individual where there was any realistic chance of that person being a civilian, without breaking the law. (In order to avoid lengthy discussion on combatant status, the rest of this paper will refer to combatants, members of organised armed groups, and civilians directly participating in hostilities as 'lawful targets'.)

Proportionality refers to a prohibition on attacks which create excessive civilian harm when compared to the military advantage to be gained.[26] It is seen as a link, and a balancing act, between military

---

[17] T. Urban, 'The AI Revolution: The Road to Superintelligence' (*Wait But Why*, 22 January 2015) <http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html> accessed 4 January 2017

[18] For opposing views on benefits and risks of AI in the future see N. Bostrom, *Superintelligence: Paths, dangers, strategies* (Oxford University Press 2016), and R. Kurzweil, *The singularity is near: When humans transcend biology* (Gerald Duckworth & Co 2006)

[19] D. Lamothe, 'Meet The Impressive Guns Protecting U.S. Bases From Rocket Attacks In Afghanistan' (Washington Post, 2015) <https://www.washingtonpost.com/news/checkpoint/wp/2015/10/21/meet-the-massive-guns-protecting-u-s-bases-from-rocket-attacks-in-afghanistan/?utm_term=.b3a1dc0d8c5a> accessed 28 September 2017.

[20] For more information see: E. Crawford and A. Pert, *International humanitarian law* (Cambridge University Press 2015), 163-194; W.H. Boothby, *The Law of Targeting* (Oxford University Press 2012)

[21] Art.48, Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), Hereafter: API

[22] Art.51(3), API. For more information on this, see: N. Melzer, *Interpretive Guidance on the notion of direct participation in hostilities under international law* (ICRC 2009); M. Schmitt, 'The interpretive guidance on the notion of direct participation in hostilities: a critical analysis' (2010) 1 Harvard National Security Journal 5–44

[23] Art.41, API

[24] Art.51(5)(a), API

[25] Art.50(1), API

[26] Art.51(5)(b) API

necessity (actions that are required to help defeat the enemy), and humanity (the avoidance of harm unnecessary for legitimate military purposes).[27] Attacks which are expected to be disproportionate are prohibited, and cannot be launched.[28] If an attack becomes disproportionate whilst it is ongoing, it must be cancelled or suspended.[29]

The principle of precautions in attack requires that during the planning of attacks, constant care shall be taken to spare civilian life and objects,[30] which requires military planners to do everything feasible to prevent civilian losses.[31] This includes choosing to attack objects which are likely to cause less civilian damage if there is a choice of attacks offering similar levels of military advantage.[32]

These principles can currently only be applied by human beings. The most advanced weaponry available today do not possess levels of cognition which would allow them to compute the relevant factors required to make LoAC decisions about all three principles. This does not, however, mean that systems in the future could not operate at the required levels of cognition to apply LoAC principles.

5. Paradigms for analysing autonomy

Currently, the two main lenses for viewing autonomy are: the 'levels' paradigm, which considers the level of autonomy that a machine has; the 'loop' paradigm, which looks at how much human oversight there is in lethal decision-making. This paper will build upon these by incorporating ideas from the ICRC, and Scharre and Horowitz. Both have analysed AWS by considering the tasks/roles which are delegate to them.[33] Adding to this concept, we will look at how legal requirements for these roles can be met when the capability of the machine performing the task is sufficient to meet the requirements.[34]

Crootof explains the 'levels' viewpoint well and considers increasing autonomy through a decreasing amount of human involvement. Her categories are: inert weapons, such as knives and conventional guns, which require human operation; automated weapons, such as landmines, which are exclusively reactive to human interference (a trip-wire sentry gun, for example, has no 'choice' in whether to fire); semi-autonomous weapons are those which may have either an ability to select or engage a target, with a human operator performing the other function (for example a semi-autonomous system may select potential targets to be confirmed and neutralised by the human operator); finally, autonomous systems, which *'are capable of selecting and engaging targets based on conclusions derived from gathered information and preprogrammed constraints, without any contemporaneous decisional support by a human being.*'[35]

Whilst the classification of weapon systems by the amount of autonomy is useful for understanding what a weapon system is capable of, it does not contain legal considerations. However, as Crootof notes herself, the level of autonomy a system has is a separate consideration to the role and tasks it carries out.[36] The present paper seeks to build in this area.

The other main analytical perspective used to investigate issues with AWS is the 'loop' paradigm. This is offered by Human Rights Watch. The three categories used are:

> '*Human-in-the-Loop Weapons: Robots that can select targets and deliver force only with a human command;*

[27] UK Ministry of Defence, *Manual of the law of armed conflict* (Oxford University Press 2004), 2.4, 2.6.2
[28] Art.57(2)(a)(iii), API
[29] Art.57(2)(b), API
[30] Art.57(1), API
[31] Art.57(2)(a), API
[32] Art.57(3), API
[33] See ICRC, n.15 above, 12-18 ; P. Scharre and M.C. Horowitz, 'An Introduction To Autonomy In Weapon Systems' (Center for New American Security 2015).7
[34] For additional thoughts on different dimensions of autonomy, see Maziar Homayounnejad in this publication, 54
[35] R. Crootof, n.12 above, 1864-1865
[36] R. Crootof, n.12 above,, 1865

Human-on-the-Loop Weapons: Robots that can select targets and deliver force under the oversight of a human operator who can override the robots' actions; and

Human-out-of-the-Loop Weapons: Robots that are capable of selecting targets and delivering force without any human input or interaction.'[37]

This paradigm firstly assumes that all AWS would be able to '*select targets and deliver force*', it does not consider weapons with lesser levels of autonomy or capability, such as systems that are purely for target identification without a component to deliver lethal force. Although, this paradigm does enable greater understanding of what it may mean to deploy an AWS with or without human control, it is a general definition enabling discussion on ethics, policy, and law. It does not directly assist in identifying specific issues for complying with LoAC. However, this is understandable, as Human Rights Watch are campaigning to have AWS banned as part of the Campaign to Stop Killer Robots,[38] which offers multi-dimensional arguments as to why AWS should be banned rather than just legal discussion.

Whilst both paradigms enable greater understanding of the issues, only the levels of autonomy paradigm offers context about what an AWS might actually be used for and neither incorporate legal issues directly. In papers by the ICRC, and Scharre and Horowitz, both consider that the type of functions which are delegated to machines are relevant to the issues which they create.[39] In this paper, we will focus on the autonomous functions of AWS related to LoAC compliance and lethal decision-making, rather than autonomous navigation for example.

Discussing the role that a system is used for enables us to build on these paradigms to identify and then evaluate more specific issues. Considering the capability of a system allows us to determine whether the system can comply with the legal requirements, and whether human assistance is needed. In this paper, we will discuss legal issues, but the questions could be adapted to enable policy and ethical discussions.

In considering the proposed role and capability which a system could be used for, we can move beyond the broad ideas that certain autonomous characteristics, or certain levels of human involvement in operations create legal issues by default. By determining particular roles and tasks which a system will be performing during the intended deployment, we can identify both the level of autonomy required for the system to perform the task, and potential legal issues. We can then discuss whether the system has the required capability to fulfil the associated legal requirements. This will then inform whether human assistance is required in each case to meet legal standards. In order to create such discussion, we need to ask additional questions when using these paradigms of analysis. This paper suggests the following additional questions:

1. What is the role of the weapon with autonomy?
2. How much autonomy does the weapon system have in this role?
3. What is/are the relevant legal issue(s)?
4. Does the weapon system have the capability to fulfil legal requirements?
5. Are there capability gaps that prevent compliance with LoAC and policy requirements? Do they need humans to fill them, and/or place restrictions on AWS behaviour in place?

These questions have been chosen to enable us to go beyond the current paradigms in common usage and identify the specific issues we are interested in for this paper. By focussing specifically on the legal requirements which will be placed upon AWS, we can evaluate whether they fulfil these necessary principles. The questions will also allow this paper to show that the relevance of certain legal issues is role-dependent. They will also allow us to show that fulfilment of these requirements is capability-dependent. Therefore, this enables us to move beyond ideas of whether a particular system will be lawful or unlawful based upon broad categories. We can make a more specific assessment based upon the actual system in use, or intended to be used, and the situation it will be used in. The more detailed analysis these questions enable will provide us with more accurate deductions about whether systems will be legally compliant beyond broad categories.

---

[37] Human Rights Watch, n.5 above, 2

[38] Campaign to Stop Killer Robots, 'The Problem' (2015) <http://www.stopkillerrobots.org/the-problem/> accessed 15 January 2017; Human Rights Watch, n.5 above, 36

[39] ICRC, n.15 above, 12-18 ; P. Scharre and M.C. Horowitz, n.34 above.7

6. Role-dependent legal issues

First, we will discuss how the relevant legal issues present are role-dependent depending upon the operation which an AWS is used for. Let us consider the task carried out by the Phalanx Close-in Weapon System mentioned above, to destroy incoming ordnance.[40] As previously mentioned, this would be classified as an autonomous system as it selects and engages its own targets without human assistance, ergo humans are off-the-loop.

Additionally, let us suppose that in the future a highly capable autonomous weapon system was to be deployed in the same role that the Phalanx currently does. As the role each system would play in this operation is identical, the legal issues raised would be the same. Despite the future-AWS having a far greater technological capability, the roles and therefore the legal issues that need to be overcome are the same as the Phalanx. Although the future-AWS would indeed have a greater level of complexity compared with the Phalanx, both would be classified as autonomous and with humans off-the-loop. This emphasises that neither of the commonly used paradigms enable discussion beyond broad categories. We will now evaluate the Phalanx and our future-AWS using our analytical questions:

Q1. What is the role of the weapon with autonomy in an operation?

The role of the Phalanx is to destroy incoming ordnance. As these munitions will all come from the enemy, the Phalanx does not need the capability to distinguish such ordnance from other potentially civilian objects, as no civilian objects could have similar characteristics to incoming rockets, artillery, or missiles within its sensor field. The Phalanx has not been designed to carry out proportionality or precautionary decisions, which have always been intended to be performed by humans on the operational level.

This role would be the same for our future-AWS in terms of being a counter-ordnance system, but could be expanded due to greater capabilities. For example, if containing the capability to perform proportionality and precautions decisions, the future-AWS could move around a military base to counter threats from different locations and perform the necessary legal decision-making unilaterally.

Q2. How much autonomy does the weapon system have in this role?

Current systems have the autonomy to choose whether an object moving within its sensor field is a target it should destroy, by comparing its sensor data with targets in its database,[41] this author assumes the Phalanx functions in the same way. Again, this would be the same for our future-AWS, but could be expanded due to greater capabilities. For example, if the capability were present, a future system could add targets to its database through machine learning.

Q3. What is/are the relevant legal issue(s)?

Clearly, distinction is of foremost importance here as the main task of the Phalanx is to identify objects as either civilian or enemy. In following the role of the Phalanx, distinction would also be the primary concern for our future-AWS.

Q4. Does the weapon system have the capability to fulfil legal requirements?

Yes, the Phalanx system has demonstrated its ability to destroy incoming ordnance on operations over the last few years.[42] Therefore, we can say that the Phalanx fulfils the role it was designed for, and the associated legal requirements i.e. distinction. However, there is no ability for the Phalanx to carry out proportionality or precautionary decisions, although it was not designed to do so. Here, our future-AWS also has the capability to fulfil the distinction requirement. With additional technology, it may also be able to perform other tasks. But, that is not relevant here.

Q5. Are there capability gaps that prevent compliance with LoAC and policy requirements? Are humans needed to fill capability gaps, or restrict AWS behaviours?

[40] Raytheon, n.13 above
[41] B. Handy, n.15 above, 87
[42] D. Lamothe, n.20 above

As the Phalanx system can recognise and destroy ordnance with reliability, there are no gaps with this system in the role of identifying and destroying ordnance. On the operational level, humans are needed to perform proportionality and precautionary decisions, as well as maintenance. Only the specific tasks involved in the counter-ordnance role can be delegated to the system. Therefore, here is a clear example of a current system carrying out a specific role within operations, and human beings filling in the capability gaps which the technology cannot yet perform.

Our future-AWS may be able to perform the wider operational tasks as well as countering incoming ordnance. Still, if a future system were to be used for the same task as the Phalanx, the legal issues would be the same: Questions would still be asked over whether it can reliably identify targets, and whether human beings are needed to fill capability gaps. The fact that our future-AWS has more capability and could potentially perform all operational tasks, does not alter the legal requirements for this individual task. So here we can summarise that the legal issues of using a particular weapon system depend upon the role in which it is being employed.

7. Capability-dependent legal compliance

Having looked at role, we will now focus on the capability of systems. Here, we will consider static border defence operations. The Samsung SGR-A1 is a South Korean sentry gun deployed along the North Korea-South Korea de-militarised zone, or DMZ. [43] This is a buffer zone where nobody from either side enters. As nobody is supposed to enter the DMZ, it could be assumed by South Korea that anybody who does actually enter the DMZ would be invading North Korean infantry. The SGR-A1 can also detect human beings present in the DMZ, and identify them as a potential target. Currently, a human being is required to authorise any weapons firing. [44] Again, as with the Phalanx system, the proportionality and precautionary decisions are taken by humans prior to use on the operational level. In terms of the existing paradigms, the SGR-A1 would be a semi-autonomous system as it can select targets, but not engage them without human authorisation; such a configuration of tasks would put the human operators 'in-the-loop'.[45] To add to this understanding, we shall ask the same questions, and have our future-AWS hypothetically perform the same role as the SGR-A1. In this scenario, the future-AWS can recognise and classify different types of people into 'unlawful target' and 'lawful target' categories, and launch attacks accordingly.

Q1. What is the role of the weapon with autonomy in an operation?

The SGR-A1 is primarily designed for identifying invading North Korean infantry. However, implicit in this is also the requirement to identify people not to target, such as civilians, medical and religious personnel, and those enemies who are sick, injured or surrendering, all of whom it would be unlawful to target. Although none of these people should be present in the DMZ, it is possible as North Korean civilians could attempt to defect to the South, and if an invasion were to happen, soldiers could be injured or attempt to surrender. As with the Phalanx, the system has not been designed to carry out proportionality or precautionary decisions, they have always been intended to be performed by humans using this system.

The role of the future-AWS would almost be the same. The role of border defence is important, and unlikely to be abandoned because a system can perform additional tasks. It may be possible for future system to carry out border defence as one of many roles, such as intelligence gathering. However, in this scenario, our system has an additional role and can launch attacks at those it deems to be a lawful target.

Q2. How much autonomy does the weapon system have in this role?

Clearly, the SGR-A1 system has the freedom to identify what it recognises as humans from other objects in the DMZ. However, as it cannot differentiate between civilians (unlawful targets) and invading infantry (lawful targets), it does not have the freedom to initiate firing its weapon without human authorisation.

---

[43] J. Pike, 'Samsung Techwin SGR-A1 sentry guard robot' (*Global Security*, 7 November 2011) <http://www.globalsecurity.org/military/world/rok/sgr-a1.htm> accessed 13 January 2017
[44] J. pike, n.44 above
[45] R. Crootof, n.12 above, 1865

Our future-AWS would also have the same freedom to identify potential humans. However, it does have the capability to recognise civilians and the enemy.

Q3. What is/are the relevant legal issue(s)?

As with the previous example the primary concern in terms of legal requirements is distinction, as the most crucial task is recognising civilian or enemy presence. This is the same for both the SGR-A1 and our future-AWS in this example. Similar to the Phalanx, the SGR-A1 cannot perform any precautionary or proportionality decisions. These must be made by humans. However, unlike the Phalanx, the system is not delegated the freedom to fire of its own volition (although it does have an 'automatic mode'). [46]

Here, our hypothetical future-AWS does have the capability no only to launch attacks itself, but also to perform proportionality and precautionary decisions. If these tasks were to be added to its role, they would create a greater number of legal issues to deal with. As civilians should not be present in the DMZ, these capabilities should not be needed. However, defecting civilians could be pursued by North Korean troops, which could raise proportionality and precautionary issues. As our future-AWS contains these capabilities, these issues can be overcome.

Q4. Does the weapon system have the capability to achieve its tasks?

Reportedly, the SGR-A1 can recognise a human being with their hands raised in surrender, [47] however that does not help the other types of person who cannot be targeted under LoAC (civilians, specially protected persons, and those *hors de combat*). There is no mention in any literature of the system being able to recognise people other than in this surrendering pose, thus it does not have the capability to perform the full task of distinction independently. So, we can say that the capability of the system does not match up to the legal requirements, thereby creating a capability gap. In this case, it is filled by human beings who authorise firing at targets which the system has selected. If this role were performed by our future-AWS, it does have the capability to achieve the full distinction role (and potentially other tasks), thereby meeting the legal requirements.

Q5. Are there capability gaps that prevent compliance with LoAC and policy requirements? Are humans needed to fill capability gaps, or restrict AWS behaviours?

Clearly, the most pressing capability gap with the SGR-A1 is the inability of the system to recognise combatants from civilians and protected persons. Thus, humans must actually identify the combatant status of anybody the system locates in the DMZ. This use of human beings to fill the capability gap, and the restriction of AWS from carrying this out themselves is required to comply with LoAC.

Although the role of the SGR-A1 and our future-AWS would be the same, the future-AWS can achieve the role without human assistance, and there is no capability gap. Therefore, it can comply with the distinction requirement of LoAC on its own.

Consequently, we can say that the difference between the SGR-A1 and our future-AWS being the capabilities they each have affects the whether or not they meet the legal requirements. Thus, compliance with the law by technological systems in these scenarios is capability dependent. Indeed, it is precisely the greater capability our future-AWS would have over the SGR-A1 in this situation that enables the legal requirements to be fulfilled, rather than creating a capability gap requiring humans to fill them. We can also deduce that it is not the autonomy itself that creates unlawful actions, but the inability of the system to perform the acts required of it by LoAC (with, or without human help).

8. As complexity of role increases, so do the number of legal issues

Now, let us consider a far more complex operation that happens daily in warzones, the armed patrol. There are no current systems for carrying out armed patrols, although the Israeli Guardium can perform

[46] J. pike, n.44 above
[47] J. pike, n.44 above

which include autonomous intelligence gathering.[48] However, let us imagine our hypothetical future-AWS is being used to perform armed patrols. In this situation, the system would be autonomous under the levels paradigm. In terms of the loop paradigm, humans are could be off- or on-the-loop. This author anticipates that if highly-capable fully-autonomous systems were to be deployed and the occurrence of engagements are unpredictable, humans will generally be off-the-loop but would begin to monitor engagements as they occur, thereby 'moving onto the loop' when fighting breaks out.

Q1. What is the role of the weapon with autonomy in an operation?

The role of an armed patrol is not simply to travel around near a military base, but to recognise potential threats, collect intelligence, and to engage the enemy if it is located. So, in reality, there are multiple tasks for a system to perform in this role. As with the other examples, recognising potential threats and lawful targets would be difficult for machines. Adding to this difficulty is the complexity of modern wars which tend to involve organised armed groups, or individual fighters who often wear civilian clothing.[49] Even highly advanced future systems may struggle to identify such fighters amongst the ordinary civilian population, so may require human assistance to actually identify the combatant status of individuals present in the area of operations. If this were to happen, the human could be seen as 'moving into the loop' by taking direct control of AWS functions. This would reduce the role of the AWS to being semi-autonomous, or even remotely controlled. The legal issues would alter accordingly, with the onus for compliance moving to the human operator.

However, let us consider our future-AWS is ambushed by an organised armed group in civilian clothing whilst on patrol. The combatant status of those firing at the AWS would become known through their actions and so the distinction criteria is simple to fulfil: those firing have self-identified as lawful targets. Although the AWS would be acting in self-defence, it would still be required to abide by the proportionality and precautionary principles. So here we have a task (defending itself) where all three LoAC principles are very much part of the systems' decision-making.

Q2. How much autonomy does the weapon system have in this role?

In terms of autonomy, the future-AWS has the freedom not only to choose its own targets, but also the freedom to determine whether or not to fire at them, which includes the proportionality and precautionary decisions. Unlike the previous examples, humans would not be able to take decisions about proportionality and precautions in attack prior to the operation, as the events would not be predictable – thus the AWS would either have to be able to make such a decision, or defer to humans for guidance. Therefore, such a system would have the autonomy to either function fully in accordance with its programming, or have the autonomy to determine that a particular situation is too complex for the system itself to deal with, and request human assistance.

Q3. What is/are the relevant legal issue(s)?

In the previous two examples, the actions of the Phalanx and SGR-A1 systems with autonomy were restricted as their roles related to the specific activity of recognising targets in an environment that had already been evaluated by humans. Thus, the main legal issue was distinction. However, in this scenario, where the environment which hostilities will take place is unpredictable, proportionality and precautionary decisions would also need to be included. Further legal considerations will also be needed; for example, considerations around protection of cultural property. A more detailed legal framework considering more issues would need to be part of the AWS programming, but specific aspects are beyond the discussion in this paper.

Q4. Does the weapon system have the capability to achieve its tasks?

[48] L. Xin and D. Bin, 'The Latest Status And Development Trends Of Military Unmanned Ground Vehicles', Chinese Automation Congress (IEEE 2013).533-534

[49] Or more on the legal framework for targeting such individuals see N. Melzer and M. Schmitt, both n.23 above

In terms of distinction, although it is not impossible that future systems could recognise enemy fighters in civilian clothes (particularly if equipped with behavioural analytics technologies)[50] they may require human assistance. If a machine could not recognise the difference between lawful and unlawful targets, it would be required to treat all persons as civilians, unless they acted in such a way as to enable their recognition as a legitimate target, i.e. firing upon the AWS.[51]

Schmitt and Thurnher state that it could be theoretically possible to reduce the information required for proportionality decisions to data which an AWS could make calculations with.[52] Thus, it could be possible for an AWS itself to fulfil the legal requirements to comply with the proportionality principle itself. However, if this technology does not come to fruition, human beings would be required to perform this task. Indeed, Human Rights Watch argue that greater levels of discretion are required for the proportionality assessment which would not be possible to programme into AWS.[53] If this assessment is accurate, humans would need to perform the proportionality decisions for such systems.

In terms of precautions, Schmitt and Thurnher describe potential AWS deployments, which are LoAC compliant, where feasible precautions to protect civilians are taken by human commanders on the operational level.[54] The replacement of human operational planners by autonomous systems is not anticipated and will not be discussed here, although some sort of AI operation-coordination system is not unimaginable. If human beings retain such decisions on the operational level, there would be no capability gap created as human beings would be meeting the legal requirements. If systems are developed which would allow greater precautions on the tactical level, for example if the systems itself could comprehend changing a missile aim-point to prevent civilians being caught in a blast radius, this would only increase civilian protections. As compliance with precautions depends upon the feasibility of actions, the level of civilian protection required by LoAC would rise with the capability of such a system.[55]

Q5. Are there capability gaps that prevent compliance with LoAC and policy requirements? Are humans needed to fill capability gaps, or restrict AWS behaviours?

In this type of scenario, where it is a highly complex environment and there are lots of tasks to be carried out in order to achieve legal requirements, our system can only fulfil the distinction requirement due to the actions of those ambushing the AWS. Our other two legal requirements may be possible to fulfil using the capabilities of the AWS, or it may not be. The complexity of the situation means that a greater level of comprehension about the impact of AWS actions on the surrounding people and environment is needed by the system in order for it to be able to make the required legal decisions. Thus, the complexity of the situation increases the number of legal issues and the level of capability and tasks that would be required of an AWS in order to comprehend the factors that are relevant to LoAC decision-making.

Therefore, we can say that as the complexity of an operation increases, the number of legal issues also increases. The tasks required to overcome those issues rises in unison. The larger number of tasks requires a larger number of capabilities to fulfil them. Consequently, the more complex an operation with AWS, the more capabilities would be required in order to fulfil the legal requirements.

9. Discussion

When discussing the examples above, we have regularly considered how human beings could intervene to fill capability gaps present in AWS. Indeed, the presence of humans is essential for legal compliance in present day systems, whether at the tactical level (as with the SGR-A1), or at the operational level (as with the Phalanx).

---

[50] N. Shachtman, 'Army Tracking Plan: Drones That Never Forget A Face' (Wired.com, 2011) <https://www.wired.com/2011/09/drones-never-forget-a-face/> accessed 4 April 2017.
[51] 'Protocol Additional To The Geneva Conventions Of 12 August 1949, And Relating To The Protection Of Victims Of International Armed Conflicts (Protocol I)' (1977). Art.50
[52] M.N. Schmitt and J. Thurnher, '"Out Of The Loop": Autonomous Weapon Systems And The Law Of Armed Conflict' (2013) 4 Harvard National Security Journal.253-257
[53] Human Rights Watch, n.5 above, 33
[54] M.N. Schmitt and J. Thurnher, n.53 above.259-262
[55] Art.57(2), API

Although the use of autonomy and artificial intelligence is representative of the vast technological progress that has been made, both present systems are actually performing a very narrow task within a larger role. Human beings are still performing a great number of tasks which these systems cannot perform. Some of the legally required tasks are discussed here, but many are beyond the scope of this paper such as maintenance, reloading, and determining where to station the weapon system. In effect, present systems have been limited to only performing they are capable of. As technology progresses, the number of tasks which machines will become capable of is only set to increase. Where they can perform tasks at the level required, these tasks can be delegated to machines. The requirements related to LoAC are set out conventions and LoAC documents. However, many of the non-legal requirements will require policy considerations about when to delegate actions to machines.

However, there may be tasks where the information to consider cannot be reduced to data and algorithmic decision-making. Human Rights Watch believe that compliance with LoAC requires human qualities which could never be converted to programming and comprehended by machines.[56] However, as we have seen, there are other opinions that suggest it may be possible.[57] What is certain is that as the legal requirements may be difficult to achieve using only machines, human beings are likely to remain involved in decision-making with autonomous systems for some time. Indeed, until the point at which systems can meet the legal requirements without human assistance, human beings will be required to partake in operations using systems with autonomy.

10. Conclusion

In conclusion, this paper has explained what autonomy is, though definitions of ordinary meaning, and in the specific field of robotics, in addition to defining AWS. It has discussed how current paradigms for analysing autonomy are useful, but do not directly enable legal analysis.

The additional analytic questions proposed for investigating autonomy consider: the role a system plays in engagements; the level of autonomy needed for that role; the specific legal issues which are relevant; whether the system has the capabilities to fulfil that role; whether there are capability gaps which need to be filled by human beings to ensure LoAC compliance.

Asking such questions enables us to recognise that autonomy in weapon systems does not, in itself, create legal issues. But, the usage of weapon systems with autonomy in roles for which they are not capable of performing would. Where this occurs, human beings could fill capability gaps to ensure compliance with LoAC. As proposed roles for AWS become more complex, the number of associated legal issues also increases. Therefore, until systems with autonomy have a large suite of capabilities which can enable compliance with LoAC, and also policy requirements, human beings are likely to be needed to fill capability gaps. Thus, fully-autonomous lethal operations are unlikely to occur for some time yet.

[56] Human Rights Watch, n.5 above, 36
[57] M.N. Schmitt and J. Thurnher, n.53 above.

Legal Personhood and Autonomous Weapons

*By Migle Laukyte**

"[…] *human history to date suggests that any new technology will be exploited for military use*"

(Cave 2012, 85)

Introduction and Premises

This paper is built on the following premises:

My *first premise* is that the research and development of autonomous weapons is not going to stop: autonomous weapons, just like nuclear weapons since the 1960s, is a way to maintain military advantage,[1] something that the most powerful and developed countries want to preserve and cultivate.[2] There are numerous initiatives inviting to stop the race of military artificial intelligence. [3] Yet those authorities who actually make decisions do not express themselves, or rather continue developing AI for war: *The New York Times* as lately as on February 3, 2017 has published the news that China is developing a cruise missile system based on advanced artificial intelligence in response to US Long Range Anti-Ship Missile. This is planned to be deployed in 2018 and it uses artificial intelligence to locate and identify enemy ships and distinguish them from neutral ships which might be in the zone of battlefield.[4]

Hence the *second premise.* The development of autonomous weapons will not stop because there are many issues at stake, such as financial commitments, and many interests to consider, such as those of military lobbies. There are two sides in this discussion: on the one hand, there are philosophers, ethicists, (some) legal scholars, pacifists, policy makers, etc. and, on the other side, there are industry players, military companies, and many others who see in autonomous weaponry not only commercial success but also no moral dilemmas involved. What makes the things even more complicated is that we are dealing with military domain which is closed and many research initiatives are being kept secret and inaccessible to the public. It would probably not be completely wrong to say that we do not really know what we are talking about.

My *third premise* is that we should not forget that questions on autonomous weapons should not cloud the bigger question on the war itself: if we forget about this bigger question, the autonomous weapons issue would be reduced to the issue of the wrong tool to perform neutral activity. But killing is not neutral, nor is the war.[5]

As to the *fourth premise*, there is—and it is so often in the debates on technology in general—an enormous confusion as concerns terminology: autonomous weapons are or are not the same as *lethal*

---

* A CONEX-Marie Curie Research Fellow of the Department of Private Law at the University Carlos III of Madrid, Spain

1 As today countries surely possessing nuclear weapons are US, Russian Federation, UK, France, China, India, Pakistan and North Korea.

2 These countries all argue that they possess nuclear weapons as means of deterrence. Furthermore, we should also bear in mind that these countries also station these weapons in other (nuclear weapon-less) countries thus creating a network of implicit consent of the international community on the matter: for instance, the US keeps some of its nuclear weapons in Europe, Turkey and Canada. My argument here is that the same strategy based on deterrence argument and implicit consent could be applied to autonomous weapons too.

3 One of examples of the initiative against military artificial intelligence is the open letter of AI and robotics researchers in 2015, which argues that "Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control" (available at https://futureoflife.org/open-letter-autonomous-weapons). A similar approach is included in the ASILOMAR AI Principles (2017) where principle n. 18 reads "An arm race in lethal autonomous weapons should be avoided."

4 J. Markoff, and M. Rosenberg. 2017. China's Intelligent Weaponry Gets Smarter. The New York Times of Febr. 3, 2017 available at https://www.nytimes.com/ 2017/02/03/technology/artificial-intelligence-china-united-states.html?emc= edit_mbe_20170206&nl=morning-briefing-europe&nlid=78636560&te=1 &_r=1

5 For the contextualization of autonomous weapons, see M. Coeckelbergh, 2011. From Killer Machines to Doctrines and Swarms, or Why Ethics of Military Robotics Is not (Necessarily) About Robots. Philosophy & Technology 24 (3): 269–78., 271. Coeckelbergh notes that "the question regarding the justification of war in general deserves more attention –not only by pacifists" and that "ethicists of military robotics should not avoid the general question about the justification of war and its meaning."

autonomous weapons? For the purposes of this paper, I will assume that they are one and the same and we are discussing autonomous weapons that can kill and, in particular, can on their own make a decision to kill someone.[6]

My *fifth premise* is based on the fact that there are also good sides in the use of autonomous weapons. Indeed, many war crimes are committed by people who find themselves in the situations of high stress and behave in ways opposite to those they adopt in their normal everyday life back home: this is something that would not happen with autonomous weapons,[7] and this is why we should substantially invest in developing and programming artificial morality for these weapons. Furthermore, the argument that autonomous weapons dehumanize war and make it a sort of an online game, where the war takes place in a parallel reality, has another side. That is that sending a machine to a warzone reduce deaths and injuries and also prevents such side effects as posttraumatic stress disorder, depression, alcoholism and other disorders related to mental health of soldiers.[8]

After having looked at these premises, my main thesis is that the discussion on autonomous weapons lacks a constructive approach. In particular, what seems to be the main problems related to these weapons, namely the problem of attributing responsibility and possibility that we might not be able to control them and they can turn (or be turned) against us, are problems, which need to be solved and not only announced. This paper is dedicated to contribute in forging such solution and my argument is based on the following reasoning:

- Legal personhood is based on real autonomy of the entity, and that
- Real autonomy of weapons is not on the agenda (yet), and thus
- The autonomy of currently available weapons is not enough to argue for the legal personhood as it is attributed to natural and artificial persons, and yet,
- *A kind* of personhood could be attributed to autonomous weapons so as to deal with responsibility gap that emerges from their use.

I will argue this thesis in the remaining part of this paper, which is organized as follows: in the next section, I briefly discuss the idea of autonomy both in general and legal terms (1.1 *Autonomy: the very idea*) and, in terms, that it applies or could apply to autonomous weapons (1.2. *Autonomy in the case of autonomous weapons*). I distinguish between real and hypothetical technological autonomies arguing that the real technological autonomy has nothing to do with legal autonomy which is linked to hypothetical (thus not yet real) technological autonomy. Then in section 2, I deal with the idea of legal personhood and, in section 3, I extend the concept of legal personhood to cover autonomous weapons and work out the *specific kind* of legal personhood: it is specific because it is not the same as the legal personhood that is recognized for artificial persons, such as states or corporations, yet it shares a few basic ideas that I work out first (4. *My suggestion*). The first subsection (4.1 *Obstacles of applying legal personhood to autonomous weapons*) is dedicated to address some of the demurs that the idea of a special kind of legal personhood of autonomous weapons could give rise to. I finish my work with closing remarks and a few ideas on the directions that my future research could undertake so as to develop and improve the ideas advanced in this paper.

1. Autonomy and autonomous weapons

1.1. Autonomy: the very idea

[6] I also assume that my conception of autonomous weapons to some extend also includes semi-autonomous weapons, namely the weapons which have their targets selected by humans, but final targeting decisions are taken by these weapons themselves as suggested by Markoff and Rosenberg 2017(n.4). I will discuss this matter later in the paper.

[7] This idea was already advanced in 2007 by Arkin who argued the following: "It is not my belief that an unmanned system will be able to be perfectly ethical in the battlefield, but I am convinced that they can perform more ethically than the human soldiers are capable of" (see R.C. Arkin, Governing Lethal Behaviour: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture. 2007 Technical Report GIT-GVU-01-11. Available at: https://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf, 7).

[8] More on the effects of war on mental health of soldiers, see Seal et al. 2007 which provides with statistical data on mental disorders of more than 100 000 US veterans coming back from Iraq and Afghanistan.

In dealing with the problem of autonomous weapons, we must first of all agree upon the terminology and, in particular, on the most important term in our discussion, namely *autonomy*.[9]

Autonomy has been a subject of study of different scholars: one of the simplest ways to see it, is provided by Gladdenwho argues that autonomy is "*entity's ability to act without being controlled*,"[10] and proceeds by explaining that

> *In its fullest form, autonomy involves not only performing cognitive tasks such as setting goals and making decisions but also performing physical activities such as securing energy sources and carrying out self-repair without human intervention*.[11]

Dworkin puts it similarly arguing that

> *I am defining autonomy as the capacity to reflect upon one's motivational structure and make changes in that structure. Thus, autonomy is not simply a reflective capacity but also includes some ability to alter one's preferences and to make them effective in action. Indeed to make them effective partly because one has reflected upon them and adopted them as one's own. [12]*

In any case, he also admits the plurality of ideas on autonomy but observes that it is a purely human feature.[13] Yet, if Dworkin was right—namely, if autonomy was only a human feature—we would not discuss the autonomy of weapons. Of course, the autonomy of weapons and autonomy of people have to be differentiated. Yet the autonomy of autonomous weapons, as distant as it might be from our human autonomy, could still be sufficient to advance a possibility of (certain kind of) legal personhood for autonomous weapons because autonomy is necessary for someone or something to be seen as legal person (natural or artificial).

There is also a legal meaning to autonomy: in legal terms being autonomous means acting on one's own behalf and not being controlled or forced to act in a certain way by someone or something else. Lapidoth has classified the legal ideas on autonomy into four main groups:[14]

(a) Autonomy as "right to act upon one's own discretion in certain matters";
(b) Autonomy as independence;
(c) Autonomy as decentralization;
(d) Autonomy as exclusive political, legislative and administrative power.

The autonomy that interests me for the purposes of this paper is the autonomy (a), described as a right to act the way one chooses. And if we see autonomy as a right, then we necessarily have also to see the duty that this right gives rise to, such as for instance, bear responsibility that the exercise of this right of autonomy brings into being.

And here is where the whole debate about autonomous weapons stands: if we do admit the autonomy of the weapons, then we give these weapons a right to decide the course of their actions in certain situations. We are not sure yet whether it is a good idea and, furthermore, we do think that this idea could turn against us. In the following section, I turn to the idea of autonomy as it is understood in technical sense by those who create and build autonomous weapons and see how their technical idea of autonomy relates to the aforementioned autonomy as a right in law.

1.2. Autonomy in the case of autonomous weapons

[9] But we also have to bear in mind that most probably the agreement on the idea of autonomy will remain fragmented for long as even in case of people we do not know yet what the autonomy really is (Sparrow, n.34).
[10] M. E. Gladden, 2016. The Diffuse Intelligent Other: An Ontology of Nonlocalizable Robots as Moral and Legal Actors, 180. In Social Robots: Boundaries, Potential, Challenges ed. M. Nørskov, 177–98. Farnham: Ashgate.
[11] Gladden, n.10
[12] G. Dworkin, 1988. The Theory and Practice of Autonomy. Cambridge: Cambridge University Press.108
[13] Dworkin, n.12, 6
[14] R. Lapidoth, 1994. Autonomy: Potential and Limitations. International Journal of Group Rights 24(1): 269–90.277

We should bear in mind that autonomy in computer and other high-tech sciences does not have the same meaning that it has in law, philosophy, sociology or other social sciences. In technical terms, autonomy refers to the fact that the tool (machine, robot, vacuum cleaner, etc.) can act without human supervision or guidance, namely, it can achieve the ends without us supervising and assisting them in the process (real technological autonomy).[15]

The whole discussion on autonomous weapons is based on the threat that autonomous weapons could get the autonomy that not only concerns the procedure to achieve the ends but also the autonomy so as to choose the ends by themselves (hypothetical technological autonomy). This is where the technological discussion on autonomy becomes philosophical, ethical and legal because the aspects and shades of autonomy as moral feature becomes intertwined with purely technical aspects of it, and autonomy to one's ends (hypothetical technological autonomy) becomes autonomy as a right to act upon personal discretion (legal autonomy).[16]

So what we can see then is that between real technological autonomy and legal autonomy there is no connection, whereas only in case the hypothetical technological autonomy would stop being hypothetical and come true, would we be able to link these two autonomies ((no longer) hypothetical and legal). Yet we seldom bear in mind this difference between real and hypothetical technical autonomies. We assume that the former is a sure antecedent of the later. It might be so and it might not be so, yet mixing up these autonomies is an error.

Of course, if hypothetical technical autonomy is not going to happen, the issue is out of discussion. The perils, however, lie not that much in wasting our time speaking about something we are not even sure is going to happen (hypothetical technical autonomy) but in failing to focus on the autonomy we already have in front of us and run aground in addressing the issues it raises. Furthermore, as real technological autonomy is not related to legal autonomy, it doesn't mean that from the legal point of view the discussion is over. In the following, I will argue that real technological autonomy does not entail legal autonomy, which is a necessary feature to become a legal person in law the way we adult human beings become. Yet there is a legal precedent when questionably autonomous entity was recognized as a legal person. My aim is to argue that something similar could work out for autonomous weapons as well.

But before we start discussing this alternative view, let us focus a little bit more on real technical autonomy: according to Krishnan autonomy can be classified according to the level of human supervision or intervention:[17]

- pre-programmed autonomy, namely, that of industrial robots, which do what they are programmed to do and nothing less and nothing more: such machines are supervised and activated by humans;
- limited or supervised autonomy which enables machine to go beyond what it is programmed to do, enabling it to much bigger variety of possible actions, yet with the need of certain supervision by the human operators;[18] and

[15] The difference between different understandings of autonomy should not be seen as an obstacle but rather than enabler: what we need to do is to look at autonomy as a boundary object—"understood as an interactive object lying at the boundary between different disciplines which makes it possible for the relative research communities to relate to one another and work together in a mutually beneficial way— between social and computational sciences" (see M. Laukyte, 2013. An Interdisciplinary Approach to Multi-Agent Systems: Bridging the Gap between Law and Computer Science. Informatica e Diritto 22(1): 223–41., 223).

[16] In this regard, Sparrow (n.34) links autonomy and moral responsibility: if an agent (autonomous weapon or human being) are autonomous enough to choose their actions according to their ends which originate in themselves and reflect their experiences and their ability to reason: if an agent is autonomous in this sense, no one else but him (it) should be held responsible for his (its) actions. In this case then we would be talking no longer about autonomous weapons, which autonomy is limited to specifically designed options, but about fully realized General Artificial Intelligence which has at least human level of autonomy (and probably also about Superintelligence, see N. Bostrom, 2014. Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.).

[17] A. Krishnan, 2009. Killer Robots: Legality and Ethicality of Autonomous Weapons. Farnham: Ashgate.

[18] Sparrow (n.34, 65) advances similar idea when he argues that sometimes autonomous weapon simply stand for a weapon "that is capable of acting independently of immediate human control; typically it means that it is a 'fire and forget' system capable of determining its trajectory or pursuing its target to some limited extent."

- complete autonomy, which does not need any human input, supervision or intervention: such machines learn by themselves and use what they learnt in practice.[19]

According to this classification, currently available autonomous weapons belong to the second class of autonomous machines: weapons with limited autonomy.

Riza develops Krishnan classification so as to work out two levels of autonomy:[20] the aforementioned

- supervised autonomy, which he re-elaborates by highlighting the increasing role of artificial intelligence, which eventually influences human supervisor's decision making: this kind of autonomy enables the machine to perform some tasks in complete autonomy whereas others with the help of human beings; and
- the so-called learning autonomy which seems similar to Krishnan's complete autonomy because it enables a robot to learn and to respond to the environment without any human intervention or supervision.

Furthermore, Riza provides with a definition of autonomous weapons arguing that that such weapons are

> *robotic weapons whose level of autonomy is based on their ability to trigger a destructive mechanism, select and identify targets, move under power or by the forces of physics, navigate to their targets, and have an ability to self-repair and self-replicate.*[21]

This is the kind of autonomous weapons I will focus on in the remaining part of this paper. I will argue that this kind of autonomous weapons possess real technological autonomy not related to legal autonomy humans possess and, consequently, cannot entail human legal personhood and responsibility (liability) to which we, adult human beings, are subject to. Nevertheless, thanks to the idea of legal fiction, we have shaped the idea of legal autonomy and legal personhood so as to cover non-human entities with a similarly unclear and questionable form of autonomy, that is, corporations: my main argument is that this idea can be further shaped so as to cover autonomous weapons, too.

2. Legal personhood

The discussion on legal personhood cannot but start with a few points as concerns the very idea of legal personhood. So what is legal personhood? What does it mean to be recognized as being a person in law?

To be a person in law can mean either that you are a natural person—a human— or a legal (artificial) person. In both cases the person is recognized as a bearer of rights and duties.[22] As a natural person, a person has rights and duties, which are different from those of a legal person.[23] For a legal person, the law recognizes specific rights and duties, such as the right to enter the contracts, the right to own property, and

[19] Yet it still doesn't mean that the machine has General Artificial Intelligence, but that it is programmed in a way that human intervention (intervention of a user) is not necessary.

[20] M.S. Riza, Killing Without Heart: Limits On Robotic Warfare In An Age Of Persistent Conflict (Potomac Books 2013).17

[21] Riza, n.20

[22] For instance, this is how the term 'person' is described in the 2016 Florida Statutes, Title I "Construction of Statutes," Chapter I *Definitions*: "The word "person" includes individuals, children, firms, associations, joint adventures, partnerships, estates, trusts, business trusts, syndicates, fiduciaries, corporations, and all other groups or combinations." Similar definition is also adopted by US Code where it reads that a person is "corporations, companies, associations, firms, partnerships, societies, and joint stock companies, as well as individuals."

[23] Although we are assisting the process by which the rights which we once considered to be only human (natural person) rights are being recognized to the artificial persons (corporations), such as the right to free speech, see A.,Mentovich, A. Huq, , and M. Cerf. 2014. The Psychology of Corporate Rights. University of Chicago Law School Working Papers No. 497/2014. Available at http://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1956&context=public_law_and_legal_theory., and the case *Citizens United v. Federal Election Commission* 558 U.S. 310 (2010). Furthermore, the right to honor also belongs to legal persons, as explained in M. I. Feliu Rey, 1990. Tienen Honor las Personas Juridicas? Madrid: Tecnos.

the right to sue and be sued in a court of law.[24] Hence, we can see that persons in law are not only humans but also artificial (non-human) persons such as companies, firms, corporations, ships and other entities.

Another important aspect of legal personhood is that it is a creation of law, namely, it exists because law permits its existence. This creation is also known as legal fiction, something that exists only in law and for the purposes of law.[25] In other words, legal personhood does not exist in nature but is a legal invention, and it means law can grant legal personhood not only to corporations but also to other entities that it will consider useful and necessary to include under the umbrella of legal personhood.

Because of the aforementioned reasoning—namely that legal personhood can be granted to anything or anyone the law considers worth granting it to—there is no reason why some sort of legal personhood could not be granted to artificially intelligent entities or other artificial intelligence based realities (robots, intelligent machines, and, eventually, autonomous weapons). Indeed, if "companies are legal constructs created for the benefit of human beings—not merely of the stakeholders, but of society as a whole" (Oliver 2015, 664), the only obstacle to grant the similar treatment to artificial intelligent entities (and the like) is to show that granting such personhood is beneficial for everyone. This is what I will do in the next section as concerns autonomous weapons. For the time being, it is sufficient to note that, for instance, if we take the basic corporate rights—such as the right to enter a contract, the right to sue and be sued and the right to own property—there seems to be no conceptual problem to extend these rights to artificial intelligent entities: electronic agents are already contracting online and the other rights could be adjusted to their specificities, too.

3. How the idea of legal personhood could be applied to autonomous weapons

So why should we think about a form of personhood for autonomous weapons? What could be the benefits of that?

First of all, let us not forget that "the idea of legal personhood has been extended for convenience to recognize, animate, and 'personify' other entities without generating confusion",[26] like it was in the case of ships—"the most living of inanimate objects"—because it is only by supposing the ship to have been treated as if endowed with personality, that the arbitrary seeming peculiarities of the maritime law can be made intelligible, and on that supposition they at once become consistent and logical."[27]

Bearing in mind the increase in complexity of autonomous weapons, we cannot ignore the problem of the chain of accountability which very much relates to responsibility and liability.[28] That is why we need to establish whether the weapon itself can be held responsible and, if not, who can.[29] As today autonomous weapons cannot be held responsible and, hence, the problem becomes that of humans standing behind the machines: people who created it, people who developed it, people who manufactured it, people who manage it, people who give orders to it, etc. The real problem to consider is to establish who did what and when, and how what someone did or did not do led to the consequences that arose.

---

[24] Of course, these are not all the rights that the legal person is entitled to: other authors add right to delegate authority to agents see J. Armour, H. Hansmann, R. Kraakman, and M. Pargendler. 2017. Foundations of Corporate Law. European Corporate Governance Institute (ECGI) - Law Working Paper No. 336/2017. Available at SSRN: https://ssrn.com/abstract= 2906054.), right to borrow money and duty to pay taxes (see W. B. Gartner, and M. G. Bellamy. 2009. Creating the Enterprise. Mason: Thomson South-Western.) and many others.

[25] More about legal theory concerning legal personhood of corporations in brief, see A. Okoye, 2016. Legal Approaches and Corporate Social Responsibility: Towards a Llewellyn's Law-Jobs Approach. Abington: Routlege..

[26] E. W. Orts, 2015. Business Persons. A Legal Theory of the Firm. Oxford: Oxford University Press.47

[27] W. O. Holmes Jr., 1881 (2011). The Common Law. Ed. P. J. S. Pereira and D. M. Beltran. Toronto: University of Toronto Law School, Typographical Society. 26. Available online at http://www.general-intelligence.com/library/commonlaw. pdf.

[28] Riza, n.20.144

[29] Riza uses the term "accountability" whereas I will use the term "responsibility" which I consider to be more inclusive and appropriate for the purposes of this paper. So in this paper the accountability will correspond to responsibility. I will use term "liability" when I will refer to the legal responsibility.

I organize people involved with autonomous weapons into two main groups: people who built (developed, tested, manufactured, provided maintenance for, etc.) the autonomous weapon (Technical Group), and people who used (military officials, soldiers, …) the autonomous weapon (Military Group).[30]

According to Riza,[31] the Technical Group most probably would not be held accountable because "military has always accepted responsibility and accountability for the actions and uses of its weapons of war," and that there are neither compelling arguments to involve manufacturers nor designers of the autonomous weapon Hence, the only responsible party responsible would be a military commander (and the Military Group), whose ability to understand, to be up to date and to manage increasing autonomy of the autonomous weapons Riza himself questions.

So what should or could we do about this situation? In the remaining part of this paper I will advance a solution—an aforementioned benefit of recognizing some sort of legal personhood to autonomous weapons—that would help us to deal with "insidious nature of increasing autonomy, which tends to diffuse responsibility and may ultimately affect accountability" and prevent the situations in which "responsibility becomes so dispersed that no one or nothing is accountable."[32]

4. My suggestion

My proposal is to attach a specific kind of legal personhood to the autonomous weapons to ensure that in case of autonomous weapon's malfunctioning or any other unforeseen and/or uncontrollable misbehaviour, the lack of current forms of responsibility (and liability) would not be an obstacle to bring justice. That ought to safeguard (or warrant?) that the use of autonomous weapons does not imply impunity and arbitrariness.

Adding legal personhood to an autonomous weapon could not only add "an active and more energetic quality"[33] to legal theory of personhood but could also contribute to the new idea of responsibility (liability) that deems to be mandatory having in mind the use of more and more complex technologies in our everyday life. In particular, I argue that with autonomous technologies—including but not limited to autonomous weapons—the overall problem is not so much *to attribute* responsibility (liability) to someone or something? but rather *to distribute* it. Therefore, we should focus not so much on pointing the finger at the responsible person. In case of complex technological tools there are many people who worked on it. Establishing the responsibility of the malfunctioning might become either impossible or completely inefficient in terms of time and costs. It is more important to rather see the involved people (as well as autonomous weapon itself) as jointly responsible for the actions (or inactions) of this weapon and distribute the responsibility among all of parties involved.

Therefore, the possible solution to this problem is to distribute responsibility among the stakeholders, which are both producers (creators, manufacturers, …) and users of the technology (both Technical and Military Groups) so much so, as argued by Sparrow,[34] otherwise it might be quite possible that no one is responsible: programmers, for instance, can avail themselves arguing that they issued sufficient warnings concerning the self-learning abilities of the autonomous weapon, which could modify its behavior in unpredictable ways, whereas military people can argue similarly that they did not command to kill civilians and that the autonomous weapon made its own decision.[35]

---

[30] I classify people into these two groups on very general grounds and I am conscious that mixed groups are what happens most likely in the real world, for instance when the maintenance of autonomous weapon is entrusted to soldiers with technical background. Nevertheless, for the sake of clarity I assume that we have these two separated groups of people. Furthermore, some authors see here responsibility of political authorities as well, but I will not include them in the picture for the moment. See U. Pagallo, 2011. Robots of Just War: A Legal Perspective. Philosophy & Technology 24(3): 307–23., 316

[31] Riza, n.20.146

[32] Riza, n.20.146

[33] Orts, n.25, 40

[34] R. Sparrow, 2007. Killer Robots. Journal of Applied Philosophy 24(1): 62–77.

[35] Yet we should not forget Riza's (n.20) argument on military's inclination to assume any kind of responsibility for whatever damages, harm or deaths that the tools they use have caused. Sparrow (n.35, 74) too holds that "the only

Furthermore, although I emphasize distribution rather than attribution of responsibility, I also agree that there should be a possibility to hold someone or something responsible for the autonomous weapons' functioning: otherwise, as suggested by Sparrow, "it would be unethical to use them in war."[36] The difference of my argument with respect to Sparrow's is that I argue that by holding someone responsible we should not limit ourselves to human beings: I argue that holding an autonomous weapon responsible, in a similar way, in which we hold corporations responsible both for their civil wrongdoings and criminal actions, could be the way out of this deadlock.[37]

In particular, my suggestion is to grant an autonomous weapon a kind of legal personhood, which would recognize the weapon as a kind of legal person (quasi-legal personhood or similar) with its own capital and shareholders (namely the two aforementioned Technical and Military Groups). This new legal person of autonomous weapon would have a capital, which would be used as an fund to cover eventual damages. For the case that damages would exceed the fund, both Technical and Military Groups would respond jointly to cover the outstanding amount.[38]

In this scenario, the autonomous weapon would be no longer seen as a tool but rather as a person on its own right, which carries responsibility for what it does in cases when its behavior exceeds the foreseen limits or range of approved actions. For obvious reasons, this personhood of an autonomous weapon would fall apart in cases, in which any member of Technical or Military Group would act maliciously: in such cases the responsible person would respond separately and the personhood of the autonomous weapon would be irrelevant.

This scheme of a kind of legal personhood would be a way to deal with the unpredictability of an autonomous weapon: when it would do something, it was not supposed to do and cause harm or damages it wasn't supposed to cause, then the autonomous weapon itself would be called to answer with the funds it has (namely, with the funds the two Groups put there for it). This is a parallel with a corporation, which responds with its own resources (put there by its shareholders) for whatever its agents have done on its behalf.

Another benefit would be transparency: today, as Riza argues,[39] the Military Group assumes all the responsibility for malfunctioning of autonomous weapons, which means that all the investigation of malfunctioning causes, settlements with the injured parties and other aspects are not public but hidden by military secrecy. Highlighting the role of Technical Group would be a way to pierce the military veil, making the settlements and investigation of malfunctioning more public thus avoiding secrecy and unnecessary speculation.

Another parallel with corporation is that the autonomous weapon as legal person would be made of constituent parties (weapon itself, Technical and Military Groups) the way the corporation is made of people (or other corporations made of people). Indeed, from this perspective autonomous weapon as legal person could be seen as a transitional legal artificial person between human based corporation on the one side, and future artificial entity which would become a person in law in its own right.

Furthermore, similarly to the corporation we would see the autonomous weapon as a result of corporate and not joint action: joint action is when the members of a group work together for the same goal and as a consequence of action the entity emerges whereas in case of corporate action, the entity is formed before acting and only then the action starts. It means that only once the autonomous weapon is recognized as a legal person (or kind of it), its life as an individual entity would start, and its duties and obligations would

possible solution seems to be to *assign* responsibility … to the commanding officer who orders their [autonomous weapons] use."

[36] Sparrow, n.34

[37] Furthermore, Sparrow (n.34) compares autonomous weapons to weapons of mass destruction or anti/personnel mines arguing that all of them do not distinguish between soldiers (legitimate targets) and civilians (illegitimate targets). Yet this is not absolutely true (and will be even less with time) because autonomous weapons, differently from weapons of mass destruction or anti-personnel mines, will be able to distinguish between soldiers and civilians thanks to artificial intelligence-based tools (face recognition, etc.).

[38] Further possibility is to insure the autonomous weapon as well and consider autonomous weapon's capital as an insurance fund.

[39] Riza, n.20

come into being.[40] Additionally, similarly to the corporation, the Technical and Military Group could establish a charter or kind of by-laws to regulate their role within the autonomous weapon.

Needless to say, how approximate and immature this idea is, and yet, I believe that it could bring some clarity into the discussion. In the following section, I will address only a couple of critical points that I see in this suggestion and try to figure out the ways how we could deal with them.

4.1. Obstacles of applying legal personhood to autonomous weapons

Of course, the idea to shape the concept of legal personhood for autonomous weapons does not come without problems.

The first problem is that the corporation as a legal person expresses itself through its human representatives. In case of autonomous weapons that would be difficult to do, unless we consider the human representative in autonomous weapon's case to be a soldier (the last to have the possibility to stop the weapon) or rather we see the autonomous weapon itself as its own representative, which would require us to distinguish between autonomous weapon as legal person and autonomous weapon as its own representative.

Indeed, if the reasoning above is sound, the autonomous weapon representing itself could overcome another obstacle: corporations have human representatives because this is the only way they communicate with and express their will to their environment. An autonomous weapon, which can speak or express itself in any language (and looking at Siri and similar tools it doesn't seem too far away from becoming a reality), the human representative would not even be necessary. These representatives are only necessary when they have to "exercise language, reasoning, and legal argument on their [corporations, children, …] behalf"[41] but this would not be the case of autonomous weapons empowered with artificial intelligence and language skills.

Another problem with this idea is that it does not deal with the growing autonomy of these weapons and, in best of cases, would be a cork by which we, metaphorically speaking, halter but do not stop the flow. Yet it is still better than nothing and, if we accept this kind of legal personhood as temporal solution, it could at least give us some time to work out a better one.

5. Conclusions and Future Work

When we speak about autonomous weapons, we should ask ourselves the question whether we do not want these weapons because they dehumanize war and make killing of people much more easy, or because we are afraid that one day these weapons could be used against us. We should give an honest answer to this question, which, in the majority of cases, would probably be that the distant wars bother us to a certain limit: what we fear is that these weapons might be available to our enemies or weapons themselves might (erroneously or not) no longer distinguish enemies from masters and companions. In either case, we should not build something that can destroy us, and yet nuclear weapons are here to stay and nobody is talking about their demolition.

In this paper, I suggested to look at the possibility that autonomous weapons are here to stay: what could we do about it if that was the case? I propose a modified version of legal personhood as it applies to corporations and other non-human persons that the law recognizes: I applied the concept of legal personhood to autonomous weapons and argue that it could be a solution for those cases when autonomous weapons would realize our worse fears and behave unpredictably or erroneously, causing damages or harm to people or property.

---

[40] More about the link between corporation and artificial agent (more sophisticated than autonomous weapon though) see M. Laukyte, 2016. Artificial Agents Among Us: Should We Recognize Them as Agents Proper? Ethics and Information Technology.

[41] Orts, n.26.44

As to future work, there are many questions that urge a serious discussion and the idea of legal personhood to autonomous weapons, and the aspects of legal personhood needs to be further developed, elaborated and considered (and eventually discarded as not functional). Furthermore, my idea of a legal personhood of autonomous weapon is shaped to an individual autonomous weapon, such as a robot, but I do not know whether the same idea would work in case of autonomous weapons understood as systems, networks and swarms.[42] Would their case be different and if yes, how?[43]

Another issue that interests me is whether completely autonomous weapons based on human like artificial intelligence would change the cards on the table again: in this scenario, I would be interested to inquire whether humanitarian law could have anything to say on such weapons, and whether any sort of protection could be extended so as to cover them as well.

Further research questions regard the specific types of legal personhood: the personhood of corporation is not the same as that of a ship, and I see a lot of potential in researching these differences and applying them to artificial intelligences in general and autonomous weapons in particular.

[42] Coeckelbergh, n.3.

[43] It seems though that whether we see autonomous weapon as a single robot or as a system does not make too much difference as concerns legal personhood: in both cases we need to abandon the idea that there is a full control of actions by some individual, and also that there can be only "distribution of responsibility between one single human actor and a single artefact. The network is much larger, and its nature goes beyond the human=artefact distinction" (Coeckelbergh n.3, 274). Be this network the network of many autonomous weapons and soldier, commander, manufacturer, etc., or be this network a network build of single autonomous weapon (robot), and these people, it does not change the core of the issue.

A Note on the Sense and Scope of 'Autonomy' in Emerging Military Weapon Systems and Some Remarks on *The Terminator Dilemma*

*By Maziar Homayounnejad**

Abstract

Lethal Autonomous Weapons Systems (LAWS) are essentially weapon systems that, once activated, can select and engage targets without further human intervention. While these are neither currently fielded nor officially part of any nation's defence strategy, there is ample evidence that many States and defence contractors are currently developing LAWS for future deployment. Accordingly, this brief note will proceed in three main sections. Firstly, it offers a definition of LAWS, focusing on what it is about weapons autonomy that may call for these systems to be delineated, and subject to certain additional requirements in both IHL and arms control. Secondly, it considers what kinds of weapon systems are likely to emerge as LAWS, the unique challenges posed by them and some ongoing innovations to address these challenges. Finally, there will be a summary and a few remarks on Paul Scharre's presentation, *The Terminator Dilemma*, which was delivered at the University of Barcelona's International Workshop on the 'Sense and Scope of Autonomy in Emerging Military and Security Technologies', in February 2017.

Three themes run throughout the paper. Firstly, 'autonomy' is a term of art, which must be narrowly and specifically applied to weapon systems, if it is to be useful in a LAWS context. Secondly, the purpose for which weapons autonomy and LAWS are defined is to delineate systems that may need to be subject to a) deployment restrictions, b) stronger and additional precautions in attack (in IHL) and / or c) commonly agreed rules to promote strategic stability (in arms control). Accordingly, a third theme is to distinguish between the strategic, operational and tactical levels; and to bear in mind that machine autonomy is technically more feasible at the tactical level, while it is necessary to have extensive (deliberative) human involvement at the operational and strategic levels. Indeed, the epitome of this is precisely the setting of strategic priorities and operational parameters within which tactical autonomy will operate.

Introduction

Lethal autonomous weapons systems (LAWS) are essentially "weapon system[s] that, once activated, can *select* and *engage* targets without further intervention by a human operator"[1]. While they are not yet fielded[2], LAWS are currently in early stage development in numerous States[3], and may well be fielded within the next decade or so. For the United States in particular, weapons autonomy is an integral part of its *Third Offset Strategy*[4], which aims to bolster US conventional deterrence in the face of declining force

* PhD candidate, Dickson Poon School of Law, King's College London. Thanks to Jürgen Altmann, Michael Horowitz and Wolfgang Richter for helpful comments on an earlier draft. The views contained in this note are mine and not necessarily endorsed by the reviewers. Any errors or omissions are exclusively mine.

[1] US Department of Defense (DoD) (2012) *Directive No. 3000.09: Autonomy in Weapon Systems*, page 13. Available at: http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf. But see 'Towards a Definition of Weapons Autonomy', below.

[2] But see the currently fielded Israeli *Harpy*, at (notes and text accompanying) n. 131-134, below.

[3] See (notes and text accompanying) n. 122-124, 142-146 and 153-155, below.

[4] Launched by the then US Secretary of Defense, Chuck Hagel, in a speech at the Reagan National Defense Forum on 14 November 2014. See the Pentagon's Memorandum on this, *The Defense Innovation Initiative*, 15 November 2014. Available at: http://archive.defense.gov/pubs/OSD013411-14.pdf.

structures[5] *vis-à-vis* capable adversaries[6]. Accordingly, the development and deployment of LAWS is arguably very likely to occur in the near-term, as advancing technologies combine with a perceived sense of military necessity.

LAWS will differ from more traditional means of warfare in three principal ways, and each of these differences will give rise to a specific set of legal issues that may call for regulation. Firstly, unlike traditional manned systems or remotely-piloted systems (RPS), the human operator will be kept 'out of the loop' in the *critical functions*[7] of an autonomous weapon, thereby leaving sensory equipment and software algorithms to take lethal action against specific targets. This raises questions on the capacity of LAWS to comply with international humanitarian law (IHL)[8].

Secondly, without the biological limitations and the financial cost of having a human constantly onboard and / or controlling the system, LAWS will have relatively *greater persistence* and *endurance* on deployments, yet be *more cost-effective* to operate than comparable manned systems or RPS. This will potentially give rise to a LAWS arms race, see increasingly large deployments for extended periods of time and, ultimately, may build up military tensions between potential adversaries. Together with the entrusting of critical weapons functions to software, these arguably will increase the risk of triggering an 'accidental war'[9], and will raise specific questions relating to arms control[10].

Finally, more so than with previous weapon systems, there are *contemporaneous* and *analogous* developments in the civilian sector, involving products such as civilian drones[11], surgical robots[12] and driverless cars[13]; these will incorporate similar sensory hardware and control algorithms for autonomous operation. Thus, a larger proportion of the components needed to produce a functioning LAWS (relative to traditional manned systems or RPS) is likely to have *dual-use* applications, which may lead to lethal autonomous capabilities spreading to rogue States and non-State actors. This will raise specific questions regarding the capacity of existing non-proliferation regimes to address this challenge[14].

---

[5] C. Pellerin, 'Deputy Secretary: Third Offset Strategy Bolsters America's Military Deterrence', *DoD News*, 31 October 2016: https://www.defense.gov/News/Article/Article/991434/deputy-secretary-third-offset-strategy-bolsters-americas-military-deterrence/.

[6] The First Offset Strategy being the build-up of US nuclear forces during the 1950s; the Second Offset Strategy being the development of precision-guided munitions, stealth and intelligence, surveillance & reconnaissance capabilities from the 1970s onwards. In all cases, the intention was for the US to develop a *force multiplier* that 'offset' the numerical advantage and / or rising technical capability of adversaries, thereby securing and maintaining the ability both to win a war and to deter one. See Walton, TA. (2016) 'Securing the Third Offset Strategy: Priorities for the Next Secretary of Defense', *Joint Forces Quarterly*, No. 82, 3rd Quarter, 6.

[7] That is, to select (find, fix, track) and engage (target, engage, assess) a target. See International Committee of the Red Cross (ICRC) (2014) *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*, Expert Meeting of 26-28 March 2014, page 62. Available at: https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014. See also (notes and text accompanying) n. 51-63, below, on the significance of focusing on the critical functions.

[8] ICRC, n.7

[9] P. Scharre, (2016) 'Autonomous Weapons and Operational Risk', *CNAS Ethical Autonomy Project*, pages 13-15: http://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf.

[10] J. Altmann, (2013) 'Arms Control for Armed Uninhabited Vehicles: An Ethical Issue', *Ethics and Information Technology*, Vol. 15(2), 137.

[11] See D. Hambling, A. Rutkin, and C. Gearin, 'How Drones are Learning to Find Their Own Way in the World, *New Scientist*, 30 July 2016: https://www.newscientist.com/article/mg23130842-600-how-drones-are-learning-to-find-their-own-way-in-the-world/.

[12] See E. Strickland, 'Autonomous Robot Surgeon Bests Humans in World First', *IEEE Spectrum*, 4 May 2016: http://spectrum.ieee.org/the-human-os/robotics/medical-robots/autonomous-robot-surgeon-bests-human-surgeons-in-world-first.

[13] For various forecasts, most of which put the emergence and development of driverless cars into the 2020s, see *Driverless Car Market Watch: Forecasts*: http://www.driverless-future.com/?page_id=384.

[14] Altmann n 10, pages 143-144; E.C. Ewers, L. Fish, M.C. Horowitz, A. Sander, and P. Scharre, (2017) 'Drone Proliferation: Policy Choices for the Trump Administration', *CNAS Papers for the President Series*. Available at: https://s3.amazonaws.com/files.cnas.org/documents/CNASReport-DroneProliferation-Final.pdf.

One response to these legal challenges may be to institute a comprehensive and pre-emptive ban on the development, testing, production, deployment and use of LAWS[15]. Indeed, since May 2013, this is exactly what is being pursued at the United Nations *Convention on Certain Conventional Weapons* (*CCW*)[16], by a coalition of non-governmental organisations (NGOs) known as the *Campaign to Stop Killer Robots*[17]. As of May 2017, nineteen States have also joined the call for a pre-emptive ban[18], thus providing a sizeable minority of governmental support. Meanwhile, in July 2015 the *Future of Life Institute* published an open letter – now signed by over 3,000 artificial intelligence researchers and over 17,000 'other' signatories – calling for "a ban on offensive autonomous weapons beyond meaningful human control"[19]. The widespread media attention this received ensured that concern about 'killer robots' has, to a certain extent, mobilised the general public. However, while ban proponents have clearly and vigorously argued their position[20], there remain – at least in the view of some authors – near-insurmountable practical issues that would make both the negotiation / signing and enforcement of a ban treaty unlikely to succeed[21]. Chief amongst these are the potential military advantage and force multiplying capabilities of LAWS[22], which will very likely defeat consensus towards a ban[23]; as well as numerous difficulties with verifying compliance, in the event that a

[15] Most prominently, Human Rights Watch and the IHRL Clinic, Harvard Law School (2012) *Losing Humanity: The Case Against Killer Robots*, which, at page 46, called on States to "pre-emptively ban fully autonomous weapons…that can make the choice to use lethal force without human input". At page 47, the report also called for a prohibition on the "development, production, and use of fully autonomous weapons", as well as "reviews of technologies and components that could lead to fully autonomous weapons." Available at: https://www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf. See also the Berlin Statement (2010) and the Original and 2014 Mission Statements of the *International Committee for Robot Arms Control* at: http://icrac.net/statements/.

[16] See the various State and NGO contributions at the 2016 Meeting of Experts on LAWS: http://www.unog.ch/80256EE600585943/(httpPages)/37D51189AC4FB6E1C1257F4D004CAFB2?OpenDocument. See also the Chairperson's *Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*, ¶¶ 21-22 and 71. Available at: http://www.unog.ch/80256EDD006B8954/(httpAssets)/DDC13B243BA863E6C1257FDB00380A88/$file/ReportLAWS_2016_AdvancedVersion.pdf.

[17] See: https://www.stopkillerrobots.org/ for further.

[18] For the full list, see Campaign to Stop Killer Robots, 'Country Views on Killer Robots', 23 May 2017: http://www.stopkillerrobots.org/wp-content/uploads/2013/03/KRC_CountryViews_May2017.pdf.

[19] See 'Autonomous Weapons: An Open Letter from AI & Robotics Researchers', *Future of Life Institute*, 28 July 2015: https://futureoflife.org/open-letter-autonomous-weapons/.

[20] See, for example, Human Rights Watch and the IHRL Clinic, Harvard Law School, n. 15; (2014) *Advancing the Debate on Killer Robots: 12 Key Arguments for a Preemptive Ban on Fully Autonomous Weapons*. Available at: https://www.hrw.org/sites/default/files/related_material/Advancing%20the%20Debate_8May2014_Final.pdf; (2015) *Precedent for Preemption: The Ban on Blinding Lasers as a Model for a Killer Robots Prohibition*. Available at: https://www.hrw.org/sites/default/files/supporting_resources/robots_and_lasers_final.pdf; (2016) *Making the Case: The Dangers of Killer Robots and the Need for a Preemptive Ban*. Available at: https://www.hrw.org/sites/default/files/report_pdf/arms1216_web.pdf; T. Chengeta, (2016) 'Measuring Autonomous Weapon Systems Against International Humanitarian Law Rules', *Journal of Law and Cyber Warfare*, Vol. 5(1), 63; Altmann n. 10; W.A. Wallach, 'Terminating the Terminator: What to do About Autonomous Weapons', *Science Progress*, 29 January 2013: https://scienceprogress.org/2013/01/terminating-the-terminator-what-to-do-about-autonomous-weapons/; P. Asaro, (2012) 'On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making', *International Review of the Red Cross*, Vol. 94(886), 687; N.E. Sharkey, (2012) 'The Evitability of Autonomous Robot Warfare', *International Review of the Red Cross*, Vol. 94, No. 886, 787; and D. Akerson, 'The Illegality of Offensive Lethal Autonomy' in Saxon, D (2013) (ed.) *International Humanitarian Law and the Changing Technology of War*, Leiden: Martinus Nijhoff, 65-98.

[21] For a brief overview and a direct response to Human Rights Watch *Precedent for Preemption* n.20, see R. Crootof, 'Why the Prohibition on Permanently Blinding Lasers is Poor Precedent for a Ban on Autonomous Weapon Systems', *Lawfare*, 24 November 2015a: https://www.lawfareblog.com/why-prohibition-permanently-blinding-lasers-poor-precedent-ban-autonomous-weapon-systems.

[22] This is highlighted by the US making weapons autonomy an integral part its *Third Offset Strategy* (notes and text accompanying n. 4-6, above).

[23] Crootof, n. 21. See also S. Watts, (2016) 'Autonomous Weapons: Regulation Tolerant or Regulation Resistant?' *Temple International & Comparative Law Journal*, Vol. 30(1), 177, pages 186-187, pointing to a potential "Balkanized sector of weapons law".

ban is formally agreed[24]; though this negative view of LAWS verification is not universally held[25]. In addition, there are strong arguments pointing to the normative desirability of allowing potentially beneficial nascent technologies to develop. Chief amongst these are the opportunity to remove from the battlefield the human frailties and imperfections that often lead to erroneous targeting and even war crimes[26]; and to exploit more accurate, precise and responsive technology that could lower civilian casualties[27].

For these reasons, the following will assume that LAWS will not be banned; instead, they will very likely be developed, fielded and deployed by States, but also specifically regulated[28], or otherwise be the subject of a non-binding Law of Armed Conflict (LoAC) manual[29]. In either case, there will be a need to define 'weapons autonomy' and 'LAWS', for reasons that are explained immediately below. Accordingly, this brief note will proceed in three main sections: firstly, there is an analysis of how to define both of these key terms; secondly, there is a consideration of weapon systems likely to emerge as LAWS, the unique challenges they raise and some ongoing innovations that may address these; finally, there is a brief comment on Paul Scharre's recent presentation on autonomy at the University of Barcelona's Faculty of Law[30].

Defining 'Autonomy' in Weapon Systems; Defining LAWS

Arriving at a satisfactory definition of 'weapons autonomy' and 'LAWS' is more than merely an academic exercise. The definition has jurisdictional consequences, as it will determine precisely *which* weapon systems are subject to a LAWS regulation treaty or LoAC manual; and whether we are discussing future systems only, or are looking to include existing ones too[31]. In that regard, this note will define weapons autonomy as a concept that applies only to *future* weapon systems. Indeed, given that the purpose of LAWS law and policy discussions at the *CCW* is to bring potentially 'problematic' weapon systems under a specific governance regime (or to ban them altogether), it would arguably serve little practical use – and needlessly

[24] These mostly relate to the fact that 'autonomy' is largely rooted in intangible software code, which is difficult to inspect and easy to update, once inspection teams have left. See, for example, R. Crootof, (2015) 'The Killer Robots are Here: Legal and Policy Implications', *Cardozo Law Review*, Vol. 36(5), 1837; J.O. McGinnis, (2010) 'Accelerating AI', *Northwestern University Law Review Colloquy*, Vol. 104, 366; K. Anderson, and M. Waxman, (2013) 'Law and Ethics for Autonomous Weapon systems: Why a Ban Won't Work and How the Laws of War Can', *The Hoover Institution, Stanford University*. Available at: https://ssrn.com/abstract=2250126.

[25] For an opposing and more positive view, along with some practical verification and compliance proposals – both technical and institutional – see M. Gubrud, and J.Altmann, (2013) 'Compliance Measures for an Autonomous Weapons Convention', *ICRAC Working Paper #2*, May 2013. Available at: https://icrac.net/wp-content/uploads/2016/03/Gubrud-Altmann_Compliance-Measures-AWC_ICRAC-WP2-2.pdf.

[26] These include hunger, tiredness, hatred, fear, the instinct for revenge and a wilful disrespect for IHL. See R.C. Arkin, (2013) 'Lethal Autonomous Systems and the Plight of the Non-Combatant', *AISB Quarterly*, No. 137, 4; and M. Sassóli, (2014) 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified', *International Law Studies*, Vol. 90, 308.

[27] See M.C. Horowitz, and P. Scharre, 'Do Killer Robots Save Lives?' *Politico Magazine*, 19 November 2014: http://www.politico.com/magazine/story/2014/11/killer-robots-save-lives-113010, referring to the humanitarian impact of the development of precision-guided munitions; and M.N. Schmitt, (2013) 'Autonomous Weapons Systems and International Humanitarian Law: A Reply to the Critics', *Harvard National Security Journal Features*, suggesting that LAWS may be more likely to be equipped with non-lethal munitions; have more sophisticated sensors; and faster, more accurate controllers than would be possible with a manned system.

[28] Crootof, n. 24, discussing various options, including a sixth Additional Protocol to the current *CCW*, or a more ambitious framework convention with the scope for its own additional protocols.

[29] S. Groves, 'A Manual Adapting the Law of Armed Conflict to Lethal Autonomous Weapons Systems', *Margaret Thatcher Center for Freedom, Special Report No. 183*, 7 April 2016. Available at: http://thf-reports.s3.amazonaws.com/2016/SR183.pdf.

[30] P. Scharre, 'The Terminator Dilemma: Autonomous Weapons and the Future of War', *Presentation at the University of Barcelona International Workshop: 'Sense and Scope of Autonomy in Emerging Military and Security Technologies'*, 27 February 2017.

[31] M.C. Horowitz, (2016) 'Why Words Matter: The Real World Consequences of Defining Autonomous Weapons Systems', *Temple International & Comparative Law Journal*, Vol. 30(1), 85.

expend much regulatory effort – to catch existing systems that operate well without any specifically applicable regulation[32].

The Many Faces of 'Autonomy'

In its purest form, 'autonomy' is a broad concept, deriving from the Greek 'auto' (self) and 'nomos' (law) to mean *self-ruling* or *self-governing*[33]. This has different meanings in different disciplines, with at least five possible variants: political, metaphysical (philosophical), legal, moral[34] and technical[35]. Across these various disciplines, the precise meaning of 'autonomy' can differ markedly[36]. In a LAWS context, it is only the *technical* sense of the word that matters, at least as a starting point. However, even 'technical autonomy' is not defined in any single way, and it is possible to discern two broad but interrelated groups of definitions here: the first focuses on *the human-machine relationship*[37]; the second focuses on the system's *internal manipulation of its own capabilities*[38]. Thus, one is 'outward-looking' while the other is more 'inward-looking'. As will be seen below, 'weapons autonomy' incorporates both of these dimensions, but must also integrate at least one more aspect (the task being performed) to account for the specific military context, and the humanitarian and strategic stabilising focus of the laws that will govern these systems.

The Dimensions of 'Weapons Autonomy'

In line with the above, Scharre and Horowitz consider there are three distinct 'dimensions of autonomy'[39], which this brief note will use to frame a working definition of LAWS. Firstly, there is the **level of human control**, which concerns the relationship between the machine and its human operator, as well as task allocation between the two. This also finds reflection in the Human Rights Watch report[40], which distinguished between three categories.

- 'Human-*in*-the-Loop Weapons', where the human both selects and engages the target, albeit via the medium of the machine.
- 'Human-*on*-the-Loop Weapons', where the machine selects and engages the target, but under human oversight and with the option of manual override.
- 'Human-*out-of*-the-Loop Weapons', where the machine selects and engages the target *without* any human interaction[41].

---

[32] P. Scharre, and M.C. Horowitz, (2015) 'An Introduction to Autonomy in Weapon Systems', *CNAS Ethical Autonomy Series Working Paper*. Available at: https://s3.amazonaws.com/files.cnas.org/documents/Ethical-Autonomy-Working-Paper_021015_v02.pdf, arguing, at page 17, that it may even be inimical to the aims of IHL if an overly broad definition is adopted, which prohibits or restrict the use of precision-guided munitions.

[33] A. Krishnan, (2009) *Killer Robots: Legality and Ethicality of Autonomous Weapons, Surrey: Ashgate Publishing*, page 43.

[34] D. Gracia, (2012) 'The Many Faces of Autonomy', *Theoretical Medicine and Bioethics*, Vol. 33(1), 57, identifying the first three as firmly established categories, and the fourth as a potential new category.

[35] 'Technical autonomy' is specific to robotics and the computer sciences.

[36] For example, 'political autonomy' refers to the ability of a community to govern itself and make its own laws, free from foreign or supranational control / influence. By contrast, 'philosophical autonomy' often focuses on the ability of an agent to determine their own actions and behaviour, thereby implying an inherently ethical dimension to the term.

[37] H. Hexmoor, C. Castelfranchi, and R. Falcone, 'A Prospectus on Agent Autonomy', in H. Hexmoor, C. Castelfranchi, and R. Falcone (eds.) (2003) *Agent Autonomy*, New York: Springer, 1-10, page 3.

[38] Hexmoor et al. n.27, page 4. For a brief review of the various definitions of technical autonomy in both of these broad categories, see T. McFarland, (2016) 'Factors Shaping the Legal Implications of Increasingly Autonomous Military Systems', *International Review of the Red Cross*, No. 900, 1.

[39] Scharre and Horowitz, n. 32, pages 6-7.

[40] Human Rights Watch and the IHRL Clinic, Harvard Law School, n. 15.

[41] Human Rights Watch and the IHRL Clinic, Harvard Law School, n. 15, page 2. It should be noted that human-*out-of*-the-loop systems still require initial activation by a human.

The idea of 'full autonomy' undoubtedly gravitates towards humans being *out of* the loop, although, as will be seen below, lower forms of autonomy may well be associated with the other categories[42]. Furthermore, while the 'level of human control' accords with some of the broader definitions of technical autonomy alluded to above, *weapons* autonomy is a much narrower and more specific term. Namely, to fully capture the essence of machine autonomy in a weapons context, it is necessary to qualify the 'level of human control' with internal system features and a sense of lethality.

Accordingly, a second dimension that also needs to be considered is **machine complexity**. This refers to the 'intelligence' of the system, and it raises terms such as 'automatic', 'automated' and 'autonomous'.

- *Automatic* implies that the system exhibits simple, mechanical responses to environmental input, such as how a mechanical thermostat works. In a military context, a landmine or trip-wire would be considered 'automatic'.
- *Automated* weapons are grounded in more complex, yet predictable rules-based systems that typically operate on IF-THEN-ELSE functions; similar to a computer spreadsheet. In a military context, air and missile defence systems are considered 'automated'; or, more precisely, they can be *set to operate* in 'automated mode' in relation to selecting and engaging specific targets[43].
- *Autonomous* systems are those that "execute some kind of self-direction, self-learning *or* emergent behavior that is not directly predictable from an inspection of its code"[44].

Wagner (2016) points out that, in contradistinction to the first two categories, autonomous systems will exhibit two unique features[45]. Firstly, they will have the capacity to make "discretionary decisions"[46]; hence, they will be able to "react, independently, to a changing set of circumstances without necessitating the interference of a human operator"[47]. Secondly, autonomous systems will require no human input on which *specific target* to select and engage; with which *specific munition* to engage that target; or on the *timing of weapons release*[48].

This classification also finds reflection in the ICRC's (2014) report[49], which distinguished between 'remotely controlled' systems (tele-operated devices with some automatic features, but mostly controlled by a human pilot), 'automated' and 'autonomous' systems[50]; the latter two approximating more closely to Scharre and Horowitz's categories of the same designation.

Finally, there is the **task to be performed by the machine**. This is crucial for determining whether or not a *regulatory* response to machine autonomy is warranted, as different actions will carry different levels of risk if systems malfunction, or are poorly designed or deployed[51]. For example, both a mechanical thermostat and an anti-tank landmine have humans 'out of the loop'. But if, say, the sensors become over-sensitive, the latter may kill civilians, whereas the former will be a mere inconvenience. More recently, there are the autonomous take-off and landing capabilities of the US X-47B stealth drone, hailed as a technical

---

[42] See (note and text accompanying) n. 75-76, below.
[43] Alternatively, these systems can be set to operate with a human-in-the-loop in these critical functions. If so, the system reverts to being 'automatic', in that a human-launched missile – e.g. from the *Patriot* battery – senses the speed and trajectory of its target and automatically alters its flight path to keep in line with the target. This is similarly true of a system like the *Phalanx*, which automatically adapts the direction of its Gatling gun, but only to keep its stream of ammunition aiming towards the human-selected target.
[44] Scharre and Horowitz, n. 32, page 6 (emphasis added to highlight that these are alternatives; namely, machine learning is an option, but not a requirement to meet the 'threshold' for an autonomous system).
[45] M. Wagner, (2016) 'Autonomous Weapon Systems', *Max Planck Encyclopedia of Public International Law*. Available at: http://opil.ouplaw.com/view/10.1093/law:epil/9780199231690/law-9780199231690-e2134.
[46] Wagner, n.45, ¶ 6. On the meaning of LAWS 'discretion', see (notes and text accompanying) n. 80-99, below.
[47] Wagner, n.45, also see n.80-99 below
[48] Wagner, n.45, also see n.80-99 below; and M. Wagner, (2014) 'The Dehumanization of International Humanitarian Law: Legal, Ethical, and Political Implications of Autonomous Weapon Systems', *Vanderbilt Journal of Transnational Law*, Vol. 47, 1371, page 1383.
[49] ICRC, n. 7.
[50] ICRC, n.7, pages 62 and 64.
[51] Scharre and Horowitz, n. 32. See, especially, the case examples at (notes and text accompanying) n. 58-59, below, which illustrate the extent of risk that can occur when autonomy fails *in the critical functions*.

breakthrough when they were successfully demonstrated in 2013[52]. To a software programmer, and almost certainly under the 'machine complexity' dimension, these would be considered 'autonomous'. However, neither take-off nor landing involve any lethal engagement or damage to property, and both usually occur far away from any actual or potential contact zone; alone, they offer no warfighting capability and pose no perils for civilians or other protected persons or objects. Hence, they are unlikely to qualify as 'weapons autonomy' for IHL or arms control purposes. By contrast, the **critical functions** of *selecting* and *engaging* targets for attack – including the **secondary critical tasks** of *selecting the munition* and the *timing of weapons release* – all have very real humanitarian impacts[53], and potentially destabilising effects on international peace and security[54]. To illustrate the point, it is instructive to consider the work of the US DoD's Working Group on Autonomous Weapon Systems[55]. Prior to drafting *Directive 3000.09*[56], the Working Group analysed a selection of existing (mostly supervised or semi-autonomous) technologies and case studies of past catastrophic errors, as well as measures that could have prevented them[57]. These included:

- the wrongful shooting down of Iran Air flight 655 by the *Aegis* Combat System (on board the USS *Vincennes*) in 1988, which killed all 290 passengers and crew members on board[58]; and
- two fratricide ('friendly fire') incidents in 2003 – one against an RAF *Tornado GR4* jet and another hitting a US Navy *F/A 18 Hornet* jet. Both involved the *MIM-104 Patriot* missile battery, and together killed both pilots and the *Tornado* navigator[59].

In each case, it was found that it was in the *selection* and *engagement* of the target – namely, those functions most closely related to lethality – where human judgement and IHL both applied, yet the potential for catastrophe was the greatest; and indeed where fatal mistakes were actually made, but with no clear lines of accountability[60]. Accordingly, to maintain the humanitarian focus of IHL and the strategic stabilising focus of arms control, the relevant tasks that should warrant a (self-)regulatory response are exclusively the (primary and / or secondary) *critical functions* discussed above[61]. Again, this approach is strongly reflected

---

[52] Vinson, B 'X-47B Makes First Arrested Landing at Sea', *Navy News Services*, 10 July 2013: http://www.navy.mil/submit/display.asp?story_id=75298, emphasising the difficulty of carrier-based landings and the unprecedented nature of autonomous integrated carrier operations.

[53] For example, selecting the wrong target may violate the principle of *distinction*. Releasing a munition with an unnecessarily large blast radius, or releasing it while civilian heat signatures are still visibly dispersing, may violate the *proportionality* principle; almost certainly the obligation to take feasible *precautions in attack*. See ICRC, n. 7.

[54] For example, by misperceiving the actions of an adversary, thereby leading to an unintended and unwarranted engagement that could escalate into a 'flash war'. See Scharre, n. 9; Altmann, n. 10.

[55] This was established by the Under-Secretary of Defense for Policy, to formulate the official US definition of 'autonomous weapon system' and, more generally, to draft US policy on AWS by way of *Directive 3000.09*.

[56] US DoD, n. 1.

[57] R. Jackson, (2014) 'Autonomous Weaponry and Armed Conflict', *Panel Discussion of the American Society of Int'l Law*, 10 April 2014. Full video available at: https://www.youtube.com/watch?v=duq3DtFJtWg.

[58] G.I. Rochlin, 'Iran Air Flight 655 and the USS *Vincennes*: Complex, Large-Scale Military Systems and the Failure of Control' in T.R. La Porte, (ed.) (1991) *Social Responses to Large Technical Systems: Control or Anticipation*, Dordrecht: Springer Science + Business Media, 99-126, pointing to 'scenario fulfilment', whereby personnel under pressure carry out standard training scenario responses, and ignore contradictory information that should lead to a different outcome. In this case, 18 sailors saw an incoming aircraft and believed that it matched the scenario of an attack by a lone military jet. Despite sensory data suggesting the aircraft was a civilian plane and not on the attack, they opened fire and killed everyone on board.

[59] See J.K. Hawley, (2017) 'Patriot Wars: Automation and the Patriot Air and Missile Defense System', *CNAS Ethical Autonomy Series*. Available at: https://s3.amazonaws.com/files.cnas.org/documents/CNAS-Report-EthicalAutonomy5-PatriotWars-FINAL.pdf, evaluating the causes of these incidents, drawing lessons from them, and suggesting prudent use of automation for the future. In particular, the fratricides were caused by a complex mix of: system malfunctions; inadequate testing and evaluation; poor training of operators; poor decision-making by those same operators; and unanticipated situations in the field, which had not been uncovered during testing.

[60] Jackson, n. 57.

[61] See C. Jenks, (2016) 'The Confusion and Distraction of Full Autonomy', *Presentation at the 2016 Meeting of Experts on LAWS*, 11 April 2016, pages 3-4, advocating the same narrow focus discussed here. Available at: http://www.unog.ch/80256EDD006B8954/(httpAssets)/7197832D3E3E935AC1257F9B004E2BD0/$file/Jenks+CCW+Remarks+Final.pdf.

in the ICRC report, which itself distils this focus on the critical functions from a number of other definitions of weapons autonomy[62].

Furthermore, as Scharre and Horowitz (2015) point out, autonomy may be a broad concept in that it can generally apply to any function but, as mentioned above, 'autonomous *weapon'* is an inherently narrower term and can *only* refer to the critical functions[63]. Here, the authors are acknowledging that the whole purpose of a weapon system is to *attack* a target, be that in offence or defence. Hence, to autonomise non-critical functions, while the selection and engagement of targets remain manually operated, should preclude that autonomy – however technically advanced – from making the system an 'autonomous *weapon*' as such; autonomy in weapon systems is, without doubt, autonomy in the critical functions.

Three final points should be noted. Firstly, there is a fourth possible dimension of autonomy: 'complexity of operational environment'. This is not included in Scharre and Horowitz (2015), yet it is implicit in some existing definitions[64]. Moreover, Alwardt and Krüger (2016) have singled it out as a separate 'dimension' of *weapons* autonomy[65], while Huang et al (2007) incorporate it as an integral part of their ALFUS model for unmanned *military* systems[66]. Clearly, the operating environment will affect the level of risk to civilians and other protected persons and objects, all else being equal; thus, it should be part of the definition and, even more so, form one of the bases for *deployment restrictions* and pre-deployment *precautions*.

Secondly, while the above has conveniently divided each of the dimensions into three or so sub-categories, the reality is more complex and each one is actually a *spectrum* of infinite possibilities[67]. At the very least, there are other models that provide more than three levels, like Sheridan and Verplank (1978)[68], OSD (2005)[69], Huang et al. (2007)[70] and, more recently, Sharkey (2016)[71]. This will complicate the task of drawing objective boundaries between 'human-controlled and 'autonomous' systems, by offering more options that may conceivably fall into both categories, depending on how they are applied in practice.

Finally, the three (or four) dimensions are largely independent of each other and it is possible, for instance, to have a human totally 'out of the loop', yet with such low 'machine complexity' as to effectively negate the concept of 'autonomy'[72]: a landmine is a prime example. Furthermore, a given weapon system

---

[62] ICRC, n. 7, pages 62-64.
[63] Scharre and Horowitz, n. 32.
[64] For example, P. Lin, G.A. Bekey, and K. Abney, (2008) 'Autonomous Military Robotics: Risks, Ethics, and Design', *US Department of Navy, Office of Naval Research*, page 103, referring to the "capacity to operate in the *real-world environment*" (emphasis added). See also Wagner's definition, (text accompanying) n. 47, above.
[65] See C.Alwardt and M. Krüger, (2016) 'Autonomy of Weapon Systems, *IFSH / IFAR Food for Thought Paper*. Available at: https://ifsh.de/file-IFAR/pdf_english/IFAR_FFT_1_final.pdf.
[66] H-M. Huang, E. Messina, and J. Albus, (2007) 'Autonomy for Levels for Unmanned Systems (ALFUS) Framework – Volume II: Framework Models Version 1.0', *NIST Special Publication 1011-II-1.0*, especially pages 30-35. Available at: http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=823618.
[67] Scharre and Horowitz, n. 32.
[68] T.B. Sheridan, and W.L. Verplank, (1978) *Human and Computer Control of Undersea Teleoperators*, Man-Machines Systems Laboratory, MIT, pages 8.17-8.19, explaining the 'levels of autonomy' model, which consists of ten levels of automation, going from complete human control through to complete machine control.
[69] Office of the Secretary of Defense (OSD) (2005) *Unmanned Aircraft Systems Roadmap 2005-2030*, Washington, DC: Department of Defense, page D-10, where Figure D-5 provides ten Autonomous Capability Levels. Available at: https://fas.org/irp/program/collect/uav_roadmap2005.pdf.
[70] Huang et al. n. 66, page 20, where Figure 4 provides three Contextual Autonomous Capability categories, together divided into ten reference levels; while Figure A-1 on page 64 shows a functional migration of the model to the *Future Combat System*, with five levels of autonomy in total.
[71] N. Sharkey, 'Staying in the Loop: Human Supervisory Control of Weapons' in N. Bhuta, S. Beck, R. Geiss, H. Liu & C. Kreb (2016). "Autonomous weapons systems. Law, ethics, policy". Cambridge University Press, London, 23-38, page 28, applying Sheridan and Verplank's model in a more explicit military context, and focusing on *human-machine collaboration*, specifically for designing systems with *greater humanitarian impact*. This leads the author to suggest five discrete levels of *human supervisory control* of weapons.
[72] Scharre and Horowitz, n 32, page 7; Lin et al. n. 64, page 103, authors' n. 8.

will *simultaneously* sit somewhere along all three (four) dimensions[73]; consequently, whether it should be considered a LAWS as such will depend on its *combined* positioning along all spectrums.

Towards a Definition of 'Weapons Autonomy' and LAWS

Taken together, the above points to 'weapons autonomy' as a feature that encompasses humans being *out of the loop*; in the (primary and / or secondary) *critical functions* of a weapon system; with those functions being performed by software controllers that are able to exercise *discretion*; in a potentially *complex and unstructured environment*. With these in mind, the following definition of LAWS is proposed:

"A weapon system which, once activated, can select and engage targets without further human intervention and *usually* without any human pre-selecting those specific targets; *and*, in the process, to exercise discretion and self-direction to operate in a potentially complex and unstructured environment."

At first sight, this may seem like a highly restrictive definition which, if all elements are to be *cumulatively* met, might overly narrow the scope of weapons autonomy. As mentioned above, however, the underlying concern is to *maintain a humanitarian and strategic stabilising focus*; together with the italicised words in the definition, this arguably provides appropriate flexibility. Thus, even in a targeted strike, where the specific target *is* pre-selected by humans, residual machine 'discretion' on *choice of munition* and *timing of weapons release* should retain the autonomous nature of the deployment, thereby requiring appropriate *precautions* by deploying commanders[74].

Scharre and Horowitz (2015) provide two further definitions, which cover systems with lower levels of autonomy, and which are useful in that they contradistinguish LAWS:

- **Human-supervised autonomous weapon system** (or simply, 'supervised autonomy') is a system with most of the characteristics of a LAWS, but with a human operator monitoring its performance and able to intervene via *manual override* to halt its operation, if necessary[75]. This is essentially a human-*on*-the-loop system.
- **Semi-autonomous weapon** is a system that incorporates autonomy in one or more of its targeting functions, but once launched, engages target(s) that have been pre-selected by a human operator[76]. These weapon systems typically do not exercise 'discretion' in the secondary critical functions[77]; thus, they are typically human-*in*-the-loop systems.

Common examples of 'supervised autonomy' include air and missile defences, such as the *Aegis* Combat System and the *Patriot* missile battery, mentioned above[78]. By contrast, most 'semi-autonomous' weapons are precision-guided / homing munitions, such as those incorporating the *Joint Direct Attack Munition*[79] tailkit.

Two Clarifying Comments

Before examining weapon systems likely to emerge as LAWS, two clarifying comments should be made about the definition. Firstly, 'discretion' is not used here to refer to anything like human-level intelligence, deliberative reasoning or true free will. Indeed, the software-based controllers of a LAWS will

---

[73] Scharre and Horowitz, n.32 page 5.
[74] See the 'Two Clarifying Comments' below, on the relevance of precautions.
[75] Scharre and Horowitz, n. 32, page 16. US DoD, n. 1, adds that human override should take place "before unacceptable levels of damage occur" (page 14); thus, aiming to *limit* but not necessarily eliminate unintended engagements.
[76] Scharre and Horowitz, *n.32*
[77] Though as US DoD, n. 1 notes, at page 14, a *basic* element of target prioritisation and timing of weapons release is consistent with semi-autonomy.
[78] See (notes and text accompanying) n. 58-59, above.
[79] See: http://www.boeing.com/history/products/joint-direct-attack-munition.page.

essentially be *deterministic* tools running on special-purpose computers that, however sophisticated, remain founded on the 'stored programme' concept[80]. Instead, 'discretion' is used here in a technical sense, in that the weapon system's controllers will: collect (input) data; process it; and, in accordance with that data and pre-programmed instructions, select one or more (output) options from a range of possible outcomes[81] *that may not have been sufficiently foreseeable to deploying commanders*. Namely, the option chosen will always be a logical consequence of *ex ante* programming and sensed data, and this remains true even when machine learning comes into play[82]. Yet, as the opaqueness and learning ability of algorithms rise, and as the complexity of the battlefield increases – hence, the range of sensed data becomes less predictable – there may be the *appearance* of machine discretion. Certainly, autonomous systems will undertake *stochastic* (probability-based) reasoning, which introduces greater pre-deployment uncertainty over their precise actions in a more complex battlefield[83].

This is significant for national authorities, in that they may need to take *stronger and additional precautions* before those deployments[84], to minimise the risk of unintended engagements, be that in an IHL or arms control context. However, one potential problem – also identified by other authors[85] – is that it can be very difficult and somewhat subjective to draw a line between 'automated' and 'autonomous' systems, especially based on any notion of apparent 'discretion'. That said, recall that the aim is to identify systems, which exhibit *such complexity that commanders and other persons in authority will need to take stronger and additional precautions*. Understood in this way, it is submitted that identification of machine 'discretion' can be done through weapons testing and evaluation (T&E) processes: these can be tailored to varying levels of realism, including "complex, realistic 'fight-in fight-out' scenario[s]", with extensive data collection and evaluation[86]. This is not to say that such T&E is easy to implement or cost-effective, especially where complex weapon systems and sophisticated enemy countermeasures are concerned[87]. Moreover, "a reliable method of testing and evaluating the performance of weapon systems with advanced *autonomous* features is still to be found"[88], and this undoubtedly poses yet another challenge to distinguishing 'automated' from 'autonomous' systems[89]. Nonetheless, specific T&E criteria for assessing LAWS have been identified[90], as have minimum requirements for empirical evidence during the process[91]; these can arguably be developed

[80] McFarland, n. 38. Namely, LAWS will be 'calculating machines', where "instructions entered by a human programmer are stored in the machine's memory and drawn upon to govern its operation" (page 15).
[81] In this regard, see Crootof, n. 24, who differentiates 'autonomous' from 'automated' systems by including in her definition of LAWS that the weapon system's critical actions will be "based on conclusions derived from gathered information and preprogrammed constraints…" (page 1854).
[82] McFarland, n 38, page 16, noting that while it is not immediately obvious, even a learning machine essentially executes instructions formulated by its developer. There is of course "an extra layer of abstraction between the developer and the weapon firing", which consists of new rules and algorithmic changes that originate not in the developer's mind, but in the data on which the system was subsequently trained. This extra layer of abstraction complicates the process of matching specific (attack) outcomes to specific (developer) commands, but it does not alter the fact that both the algorithmic changes and the final act of weapons release involved the machine logically "executing instructions formulated by its developer".
[83] ICRC, . 7, page 13.
[84] The terms 'authorities' and 'precautions' are used broadly here, to refer both to commanders deploying LAWS in an IHL-relevant scenario; and ministries of defence or political decision-makers negotiating arms control agreements prior to employing LAWS.
[85] For example, Crootof, n. 24; and Scharre and Horowitz, n. 32.
[86] See Qinetiq's 'Test & Evaluation' page: https://www.qinetiq.com/services-products/weapons/Pages/test-and-evaluation.aspx.
[87] A. Backstrom, and I. Henderson, (2012) 'New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews', *International Review of the Red Cross*, Vol. 94, No. 886, 483.
[88] V. Boulanin, (2015) 'Implementing Article 36 Weapons Reviews in the Light of Increasing Autonomy in Weapon Systems', *SIPRI Insights No. 2015/1*, page 15 (emphasis added).
[89] Indeed, one of the reasons for the 2003 *Patriot* fratricides was inadequate T&E. See n. 59, above, for further.
[90] Boulanin, n. 88, pages 13-14, providing a checklist that expands existing IHL and human rights-based review standards, and applies these in a LAWS context.
[91] Boulanin, n. 88, page 14, noting documentation from: the manufacturer (on system characteristics and performance); the end-user (on the concept of use); and independent and unbiased operational, medical and technical tests.

and refined over time to delineate relevant boundaries and achieve T&E goals[92]. In addition, there is a number of ways to address the T&E cost hurdle[93], although these also raise a few practical challenges of their own[94]. Furthermore, there are elements of best practice on carrying out wider *legal* reviews of new autonomous systems[95] under Article 36 of *Additional Protocol I*[96]. Importantly for LAWS, the potentially greater and stronger precautionary obligations to which these systems may give rise is addressed in the requirement of legal review panels to recommend *appropriate restrictions on deployment and use*, where performance limitations are apparent in the critical functions[97]. Such restrictions may include limits on the timing of deployment, the operational environment, or the types of targets being attacked[98]. Finally, to address the apparent subjectivity of delineating the automated / autonomous boundary, note that a recurring theme in the legal review process is the concept of 'reasonableness'[99]. Consistent with that focus, it is submitted that the commander who finds the appearance of 'discretion' and uncertain stochastic behaviour in an otherwise deterministic system should be the 'reasonable military commander'[100] – a term with which legal review panels will no doubt be familiar – operating in test conditions that aim to simulate the same battlefield complexities and 'fog of war' as those for which the LAWS is designed[101].

The second clarifying comment concerns the role of the human. To be sure, humans will not be totally 'out-of-the-loop' in the use of any LAWS that may be fielded in the near-term. Rather, commanders will retain full control of a number of important parameters, which will be programmed into the systems. As a bare minimum, these include the following three[102].

- **Target parameters**, which define the exact *categories* (types) of targets that may be engaged, such as 'tank', 'attack helicopter' or 'armoured personnel carrier'[103].
- **Geographical boundaries**, which define the *spatial restrictions* within which LAWS may loiter and engage targets[104].

---

[92] Boulanin, n. 88, pages 14-15, providing a seven-point criteria of issues that a successful T&E process should determine.

[93] Boulanin, n. 88, page 16, suggesting that the cost hurdle may be overcome by: sharing T&E burdens between trusted allies; outsourcing it to the manufacturer in the case of 'off-the-shelf LAWS', thereby splitting the cost between all customers; or by relying on computer simulations, where possible.

[94] Boulanin, n. 88, noting that States may be reluctant to share the results of their own testing to avoid future liability issues. Also, both the general weapons capabilities and specific software code for LAWS will likely be classified for strategic and security reasons, thereby posing more challenges to the pooling of T&E efforts.

[95] Boulanin, n. 88, pages 16-17, referring to earlier timing of reviews; multidisciplinary review panels; cross-disciplinary training of personnel, to improve sharing of insights; incorporating manufacturer T&E evidence; and, importantly, defining restrictions on the use of anything that automates the critical functions.

[96] *Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts*, 8 June 1977, 1125 UNTS 3, entered into force 7 December 1978.

[97] W.H. Boothby, (2014) *Conflict Law: The Influence of New Weapons Technology, Human Rights and Emerging Actors*, The Hague: TMC Asser Press.

[98] Boulanin, n. 88, page 17.

[99] J. McClelland, (2003) 'The Review of Weapons in Accordance with Article 36 of Additional Protocol I', *International Review of the Red Cross*, Vol. 85, No. 850, 397.

[100] Office of the Prosecutor (OTP) (2000) *Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign against the Federal Republic of Yugoslavia*, ¶ 50. Available at: http://www.icty.org/en/press/final-report-prosecutor-committee-established-review-nato-bombing-campaign-against-federal.

[101] As the OTP,n.100, pointed out in the context of proportionality assessments, there will be room for argument in closes cases; yet in many other cases, reasonable military commanders faced with similar circumstances will generally agree.

[102] All three are detailed in Article 36, *Killing by Machine: Key Issues for Understanding Meaningful Human Control*, 6 April 2015. This adds, at page 4, that the three parameters should be set tightly enough to enable human control over an 'individual attack', which may potentially comprise multiple acts of violence. Available at: http://www.article36.org/wp-content/uploads/2013/06/KILLING_BY_MACHINE_6.4.15.pdf.

[103] To minimise the risk of false positive errors, more refined parameters may also be programmed, such as approximate size dimensions and enemy insignia. Also, where technology and circumstances permit, target parameters may prescribe exact models that are both *exclusive* to the enemy and *recognisable* to the deployed ATR systems; for example, *T-80* tank or *AH-64 Apache* attack helicopter.

[104] As an absolute (legal) maximum, this will go no further than the geographical limits of an armed conflict that is, for example, recognised in a public statement of the ICRC. In practice, commanders are likely to set tighter geographical

- **Temporal boundaries**, which define when LAWS operations will *begin* and *end*[105]; and which may be subdivided into *loitering* and *loitering plus target engagement*[106].

This leaves the weapon system to select and engage *specific* targets that fall within these general constraints[107]; save in the case of targeted strikes[108]. A fourth dimension of human control is to have a real-time feedback loop, or some other **capability to monitor** the system[109], which enables a commander to detect and **shut down** a LAWS that may have "gone rogue"[110]; or that will face a difficult scenario, which its control software may not be able to resolve[111].

Only within these four limiting parameters will a LAWS be capable of offering genuine military utility, by enabling commanders to remain *accountable* for its actions and *responsible* for the overall outcome of operations in their own area of command and control. Concretely, such limiting parameters enable commanders to: fulfil mission objectives; in compliance with IHL norms and mission Rules of Engagement; with full situational awareness[112]; and in pursuit of the broader strategic / political and military purpose of the operation[113].

Accordingly, humans will remain in the 'wider loop' of (strategic and operational) control, while the weapon system operates with *relative* autonomy within the (tactical) 'narrow loop'[114]. As a corollary, we see that "many key targeting decisions will…be made in *earlier phases* of the targeting cycle and at *locations further removed* from the intended strike site"[115]. Thus, deliberative human reasoning on the use and deployment of LAWS will potentially be made in a more abstract setting, with less reliable knowledge of concrete threats or specific civilian risks. Again, this may require that commanders take *greater precautionary measures* in their deployment and use of LAWS at the operational level.

Weapon Systems Likely to Emerge as LAWS

boundaries, depending on where enemy targets are believed to be; but also depending on the sophistication of systems deployed and on the nature of the battlefield. Note also Article 36's position on 'individual attack' as a key limiting principle (n. 102, above). This is to increase the predictability of LAWS actions and to make systems easier to monitor, thereby minimising the risk of false positive errors.

[105] Unless there is a reasonable chance for mid-operation (e.g. air-to-air) refuelling, the temporal boundaries may be technically limited by how much fuel a LAWS can carry; certainly, by how long the system can operate before needing maintenance. Again, however, commanders are likely to set tighter restrictions, to ensure greater operational control of systems. Also note Article 36's 'individual attack' limiting principle at n. 102, above.

[106] The 'loitering only' restriction will limit the system to intelligence, surveillance and reconnaissance (ISR) functions, which may provide essential intelligence that commanders need to review before switching the system to 'engagement' mode. This may be in place of deploying separate ISR units, before deploying a LAWS.

[107] Scharre and Horowitz, n. 32. This includes discretion over the secondary critical functions.

[108] Where, as mentioned above, human commanders pre-select the *specific* target, and machine discretion is limited to the secondary critical functions, to achieve the mission objective while minimising civilian harm.

[109] This need not be a continuous visual or audio-visual link, which would largely negate the concept of 'weapons autonomy'. Instead, the LAWS may be programmed to detect particularly unusual, ambiguous or unexpected battlefield scenarios, before contacting commanders for further instruction. See W.H. Boothby, 'Autonomous Attack – Opportunity or Spectre? in T.D. Gill, (ed.) (2015) *Yearbook of International Humanitarian Law 2013*, Vol. 16, The Hague: TMC Asser Press, 71-88, page 83.

[110] K. Egeland, (2016) 'Lethal Autonomous Weapon Systems under International Humanitarian Law', *Nordic Journal of international Law*, Vol. 85(2), 89, pages 102-103.

[111] Shutting down the system would be an extreme response; reverting to remote-piloting may also be an option.

[112] This includes the status and movements of: own and allied forces; enemy forces; and the civilian population.

[113] I am grateful to Wolfgang Richter for pointing out these linkages.

[114] AIV and CAVV 'Autonomous Weapon Systems: The Need for Meaningful Human Control' *Report No. 97 AIV / No. 26 CAVV*, October 2015. Available at: http://aiv-advies.nl/download/606cb3b1-a800-4f8a-936f-af61ac991dd0.pdf. See (notes and text accompanying) n. 207-213, below, for a concrete application of this.

[115] J.S. Thurnher, 'Means and Methods of the Future: Autonomous Systems', in P.A.L. Ducheine, M.N. Schmitt, and F.P.B. Osinga (eds.) (2016) *Targeting: The Challenges of Modern Warfare*, Hague: Asser Press, 177, page 178 (emphasis added).

An important question that remains is what kinds of real-world systems will likely fit the above definition of LAWS. To a certain extent, the answer to this will be broad and hypothetical, as currently no State is *officially* developing and fielding any specific autonomous weapon systems[116]. That said, there have been various roadmaps and projections[117]; recommendations[118]; statements from the US[119], Russia[120] and the Netherlands[121]; the testing of autonomous systems concepts by the DoD[122] and the US Navy[123]; as well as full-scale demonstrators from defence contractors[124]. All of these point to a strong likelihood that LAWS will be developed and fielded in the foreseeable future. This is not to mention the fear in the US that its two biggest adversaries, Russia and China, may be heavily investing in robotics and autonomy, in a field where "[e]arly adoption will be a key comparative advantage, while those that lag in investment will see their [military] competitiveness slip"[125].

Three Potential LAWS Categories

Accordingly, Horowitz (2016) distinguishes three probable categories of LAWS that may arise as part of such rivalry.

- **Munitions**: physical and non-returnable / inherently one-way weapons designed to destroy a single target. Manned versions of these include rifle and cannon rounds.
- **Platforms**: inherently returnable systems that launch munitions. Manned versions include combat aircraft, tanks, warships and submarines.

---

[116] See, for example, the various State contributions at the 2016 Meeting of Experts on LAWS, n. 16, in which the US noted that its key policy document, *Directive 3000.09*, merely sets out how the US *would* consider proposals to develop LAWS, though without establishing a firm position on future LAWS development; while the UK expressly ruled out any official policy of developing LAWS designated for offensive attack.

[117] For example, OSD (2005) n. 69; US DoD (2013) *Unmanned Systems Integrated Roadmap: FY 2013-2038*. Available at: https://www.defense.gov/Portals/1/Documents/pubs/DOD-USRM-2013.pdf.

[118] For example, US DoD (2016) *Report of the Defense Science Board Summer Study on Autonomy*. Available at: https://www.hsdl.org/?view&did=794641.

[119] US DoD, n. 117, stating, at page 67, that US Navy and DoD leadership have identified autonomy in unmanned systems as a "high priority", both to overcome access threats and to maximise fleet capacity.

[120] For example, Russian Chief of the General Staff Gerasimov reportedly said that in the near-future, the Russian military may have "a fully roboticized unit…capable of independently conducting military operations". See Deputy Secretary of Defense Speech (Robert Work), *CNAS Defense Forum*, 14 December 2015. Available at: http://www.defense.gov/News/Speeches/Speech-View/Article/634214/cnas-defense-forum.

[121] Government of the Netherlands (2016) *Government Response to AIV / CAVV Advisory Report No. 97, Autonomous Weapon Systems: The Need for Meaningful Human Control*, expressly agreeing with the advisory report that "…if the Dutch armed forces are to remain technologically advanced, autonomous weapons will have a role to play, now and in the future". Available at: http://aiv-advice.nl/8gr#government-responses.

[122] For example, the recent testing of a 103 micro-drone swarm, which demonstrated advanced swarm behaviours, such as collective decision-making, adaptive formation-flying and self-healing, all without human direction. See US DoD 'Department of Defense Announces Successful Micro-Drone Demonstration', *Press Release No. NR-008-17*, 9 January 2017: https://www.defense.gov/News/News-Releases/News-Release-View/Article/1044811/department-of-defense-announces-successful-micro-drone-demonstration/.

[123] For example, the US Navy's LOCUST programme recently ran successful trials of eight-drone swarms controlled by a single person penetrating sophisticated ship defences. The aim is to increase this to 30-drone, then 50-drone swarms controlled by a single operator, all at a lower cost than a single Harpoon anti-ship missile. See Hambling, D 'US Navy Plans to Fly First Drone Swarm this Summer', *DefenseTech*, 4 January 2016: https://www.defensetech.org/2016/01/04/u-s-navy-plans-to-fly-first-drone-swarm-this-summer/.

[124] For example, the British *Taranis* stealth drone demonstrator developed jointly by the UK Ministry of Defence and BAE Systems. See 'Taranis', *BAE systems Products*: http://www.baesystems.com/en/product/taranis. Specifically on *Taranis*'s autonomous capabilities, see Stevenson, B 'Analysis: Taranis Developers Reveal Test Flight Specifics', *FlightGlobal*, 16 May 2016: https://www.flightglobal.com/news/articles/analysis-taranis-developers-reveal-test-flight-spec-425347/.

[125] Work, n. 120.

- **Operational system**: a battle planning tool that defines mission goals, selects targets, allocates resources to achieve those targets and executes combat operations[126].

In the near-term, the kind of weapon that can be expected to emerge as an 'autonomous' system is the 'wide-area search-and-attack' loitering *munition* and *drone*[127]; operating standalone[128] or, increasingly, as part of a swarm[129], where collective behaviours can bring yet more dramatic and disruptive change to military operations[130]. Presently, a rather rudimentary autonomous **munition** exists in the Israeli *Harpy*. This detects and engages specific radar-emitting objects, with the option of negative visual confirmation[131], and all within tight spatial and temporal boundaries within which deploying commanders believe lawful targets exist[132]. That said, the *Harpy* consists of many of the same technologies in today's semi-autonomous homing munitions, but with greater range and aerial persistence[133]; thus, it is mostly 'LAWS by usage'[134]. Accordingly, we can expect future incarnations to build on this: to be more sophisticated by incorporating stronger artificial intelligence (AI) and automatic target recognition (ATR) capabilities; and to have longer loitering times and greater loitering areas. This will enable those munitions to engage a wider range of targets with improved target accuracy; for example, by loitering longer for more cross-cueing opportunities. It will also endow them with the intelligence to detect, recognise and mitigate some civilian risk; for example, by varying the exact timing of attack in accordance with circumstances on the ground.

Perhaps even before this occurs, *swarming* munitions are more likely to be developed, fielded and deployed first[135]. Generally, swarms can be utilised in three ways: *attack*; *defence*; or in a *support* role, like providing intelligence, surveillance and reconnaissance (ISR)[136]. In all cases, they involve "large numbers of dispersed individuals or small groups coordinating together and fighting as a coherent whole"[137]. Within these, the *individual* agents follow simple rules, from which the swarm *collectively* exhibits emergent intelligence, and complex and unified behaviours[138]. Human controllers supervise missions at the *operational* level, while individual autonomous units manoeuvre and perform various tasks unaided at the *tactical* level[139]. Accordingly, swarms may tend closer towards human-*on*-the-loop systems[140], although this is not necessarily the case as they may also be deployed in fully autonomous mode[141]. An example of swarming – albeit in an ISR context – is the recent testing of a swarm of 103 Perdix drones, which demonstrated collective decision-making, adaptive formation flying and self-healing[142], all with a human

---

126 Horowitz, n. 31, pages 94-97. Specifically on the nature of an operational battle planning system, see Corn, GS. and Corn, GP. (2012) 'The Law of Operational Targeting: Viewing the LOAC Through an Operational Lens', *Texas International Law Journal*, Vol. 47(2), 337.
127 Scharre and Horowitz, n. 32; Horowitz, n. 31.
128 Scharre and Horowitz, n. 32; Horowitz, n. 31.
129 I. Lachow, (2017) 'The Upside and Downside of Swarming Drones', *Bulletin of the Atomic Scientists*, Vol. 73(2), 96.
130 P.. Scharre, (2014) *Robotics on the Battlefield Part II: The Coming Swarm*, Washington, DC: CNAS, discussing how swarms of robotic systems can bring greater mass, coordination, intelligence and speed to the battlefield, thereby increasing the chance of gaining a decisive advantage over adversaries. Available at: https://s3.amazonaws.com/files.cnas.org/documents/CNAS_TheComingSwarm_Scharre.pdf.
131 Namely, there is the option to visually zoom-in on the radar-emitting object, compare the image with a database of known 'friendly' sites; if no match is recognised, the *Harpy* proceeds to dive-bomb into its target.
132 P. Scharre, 'Autonomy, "Killer Robots," and Human Control in the Use of Force – Part I', *Just Security Blog*, 9 July 2014b: https://www.justsecurity.org/12708/autonomy-killer-robots-human-control-force-part/.
133 Scharre*, n.132*
134 Horowitz, n. 31 points out, at pages 92-94, that the operator's *usage* can make semi-autonomous weapons fully autonomous.
135 See, more generally, D. Hambling, (2015) *Swarm Troopers: How Small Drones Will Conquer the World*, London: Archangel Ink.
136 Scharre, n. 130.
137 Scharre, n. 130, page 26.
138 Scharre, n. 130, page 26.
139 Scharre, n. 130, page 26.
140 See (note and text accompanying) n. 41, above, on the Human Rights Watch classification of autonomy.
141 At least in the near- to mid-term, the fully autonomous deployment option may be necessary in times of symmetric (high-intensity) conflict and / or where loss of communication links may be expected.
142 See n. 122, above.

operator defining broad tasks but not instructing any of these specific behaviours. The US Navy's LOCUST programme (Low-Cost UAV Swarming Technology) goes a step further and aims to utilise swarming munitions for ship defences[143]; while those designed for offensive attack will have the advantage of saturating and overwhelming enemy defences, such that 'leakers' still get through and take out their target[144].

Platform-based LAWS can be expected to be a step up from the loitering munition, enabling even *longer loitering* and a *choice of munitions* in attack. This may include a variety of different blast radiuses, which the control software may be able to match to its immediate environment before weapons release (to further mitigate civilian risk); and also non-lethal munitions, to warn civilians to flee and / or incapacitate combatants rather than kill them, should this be consistent with mission goals. Again, swarming is very likely here, the US Navy having already successfully tested autonomous swarm boats, both defensively and for offensive attack on hostile vessels[145], with future applications to aerial vehicles expected[146].

By contrast, autonomous **operational systems** are "the most akin to science fiction"[147], as they would call for a high degree of deliberative reasoning and strategic thinking, which is unlikely to be viably automated in the near-term[148].

The Unique Challenges Posed by the Different LAWS Categories

After pinning down what the different types of LAWS might do, a crucial and related task is to figure out "the unique challenges those [L]AWS might raise for the lawful use of force"[149]. In that regard, Horowitz (2016) argues that autonomous **munitions** are likely to pose little more challenge for IHL compliance than today's semi-autonomous munitions, the main issues being *when*, *where* and *how* such systems are deployed at the operational level[150]. In reality, there are likely to be added complications where longer loitering times, wider loitering areas and more generalised target parameters are at issue; nonetheless, these challenges should all lend themselves to existing solutions for ensuring the lawful use of force, such as the effective training of personnel and elaborate battle planning, which incorporates a wide range of precautionary options for commanders to consider[151]. Autonomous **platforms** will raise yet more potential challenges, because of their greater aerial persistence, which brings even more 'human out of the loop' time in which unintended engagements might occur. Furthermore, multi-target engagement may lengthen chains of causation and bring more unforeseeable consequences. However, these may not be insurmountable challenges, and it is again possible that clear rules and training on *when*, *where* and *how* to deploy autonomous platforms at the operational level, along with detailed battle planning, will minimise the operational risk[152].

It should be noted that while early LAWS models – particularly standalone units – may be primarily intended for deployment in traditional (sparsely populated) battlefields, research plans are currently underway at DARPA[153] to develop and fine-tune drone swarming *tactics* for urban environments[154]; this

---

[143] See n. 123, above.
[144] Scharre, n.130.
[145] D. Smalley, 'The Future is Now: Navy's Autonomous Swarm Boats can Overwhelm Adversaries', *Office of Naval Research News & Media Center*, 5 October 2014, demonstrating up to 13 boat swarms armed with .50 calibre machine guns, escorting a high-value Navy ship before detecting an enemy vessel and swarming around it: https://www.onr.navy.mil/en/Media-Center/Press-Releases/2014/autonomous-swarm-boat-unmanned-caracas.
[146] Smalley, n.145
[147] Horowitz, n. 31, page 96.
[148] Horowitz, n. 31, page 96.
[149] Horowitz, n. 31, page 94.
[150] Horowitz, n. 31, page 95.
[151] Corn and Corn, n. 126; also, G.S. Corn, and J.R. Schoettler, (2015) 'Targeting and Civilian Risk Mitigation: The Essential Role of Precautionary Measures', *Military Law Review*, Vol. 223(4), 785.
[152] Horowitz, n. 31, page 96.
[153] That is, the Defense Advanced Research Projects Agency. See: http://www.darpa.mil/.
[154] See DARPA, 'OFFSET Envisions Swarm Capabilities for Small Urban Ground Units', *DARPA News and Events*, 7 December 2016. Available at: http://www.darpa.mil/news-events/2016-12-07.

applies to both munitions and platforms. The goal is to develop new *uses* (tactics) for *existing* drone technologies, especially to support ground troops in urban conflict zones; in this sense, the DARPA project may also be 'LAWS by usage'. As mentioned above, swarms often exhibit emergent behaviours, but where they operate with a 'human-*on*-the-loop', unlawful behaviours can usually be manually overridden. That said, with the higher tempo of combat and the potentially large numbers of tactical units under the supervision of a single operator, there is still a need for new command and control paradigms to enable humans to deploy large swarms effectively[155].

Finally, autonomous **operational systems** will, far and away, be the most legally problematic category. The deliberative nature of the reasoning and the extensive number of finely-graded options that often arise in high-level combat planning means that the observed behaviour of the system would become even more unforeseeable by those deploying it[156]. Concretely, longer chains of causation and a greater disconnect between responsible persons and the actual selection and engagement of targets[157] all mean that – barring any major technological shift – such systems will likely be overseen by humans, if not remaining mere tools in the near-term; namely, decision-support systems in an essentially human-centred planning process.

Some Ongoing Innovations to Enhance the Performance and IHL Compliance of LAWS

In the near-term, there is a range of new or enhanced scientific applications – many of which involve quantum physics[158] – that may increase the range and quality of sensory data available to LAWS. These can potentially mitigate some of the challenges identified above, and they include some of the following.

- 'Ghost imaging', which is a technique still being developed and refined at the *US Army Research Laboratory*[159]. This will increase the likelihood of accurately detecting and perceiving military targets, thereby avoiding inadvertent attacks on civilians or other protected persons or objects[160].
- Devices like smaller, more accurate and penetrating gravimeters – again, an ongoing research project – which can enable the monitoring of underground and undersea movements[161]. This has a host of military applications that may enhance target identification[162], as well improving collateral damage estimation methodology[163].

---

[155] Scharre, n. 130, pages 38-41, discussing centralised coordination, hierarchical control, coordination by consensus, and emergent coordination as potential command and control models. It should be noted, however, that these are developing areas of research and it is not yet clear that they will be fully IHL-compliant.
[156] Horowitz, n.31, page 96.
[157] Horowitz, n.31, page 96.
[158] See, more generally, *The Economist Technology Quarterly: Q1 2017*, 9 March 2017: http://www.economist.com/technology-quarterly/2017-03-09/quantum-devices.
[159] 'The Newest Thing in Quantum Imaging', *Department of Defense Armed with Science*, 3 January 2014: http://science.dodlive.mil/2014/01/03/the-newest-thing-in-quantum-imaging/.
[160] *Economist Technology Quarterly* (2017) n. 158, explaining that ghost imaging works by combining pictures of a target object (along with all the heat- and smoke-based distortions generated by military action) with light beams reflected directly from the target. Correlating those measurements derives an artificially-generated, but vastly improved holographic image of an object that might be two or so miles away on a smoky battlefield. Essentially, the system is computing the paths that light takes to the target and back to the sensors, and it corrects for distortions on the actual image.
[161] S. Perkins, 'Tiny Gravity Sensor Could Detect Drug Tunnels', *Science*, 30 March 2016: http://www.sciencemag.org/news/2016/03/tiny-gravity-sensor-could-detect-drug-tunnels-mineral-deposits, describing the development towards miniaturised gravimeters that can be installed on drones to detect manmade tunnels used for drug smuggling, and also to detect underground chemicals or mineral deposits.
[162] On land, improved gravimeters may enable a LAWS to: track enemy movements through underground tunnels; to continue pursuing a target that disappears into a tunnel, despite visible heat signatures ceasing; and they may also help to detect the movement and storage of military supplies underground. At sea, improved gravimeters will be better able to spot moving masses underwater, such as submarines and torpedoes.
[163] For example, by detecting underground objects and chemicals that may increase the collateral effects radius of a given attack. This will enable a LAWS to make more accurately informed attack decisions.

- These same devices will also enable the safe navigation of LAWS in GPS-denied environments[164], which is particularly important in symmetric (high-intensity) conflict.
- Recent advances in electro-optical (vision-based) guidance systems are also enhancing targeting accuracy, while enabling GPS-denied operation and eliminating the risk of GPS hacking[165]. One particular application of this, Raytheon's *Multi-Spectral Targeting System*, is now available in miniaturised form[166], which can only increase its applicability to a wider range of LAWS.

As mentioned above, both the intended operational environment and the standard of technology available will affect such legal matters as *deployment restrictions* and the level of *precautions* to be taken by deploying commanders. Arguably, these innovations will have a positive effect on deployment options, while lowering the necessary level of precautions required for safe use.

To conclude, wide-area search-and-attack loitering *munitions* and drone *platforms* are likely to be the first LAWS developed and fielded, with swarms potentially coming sooner than standalone units – assuming, of course, that no pre-emptive ban is put in place. Furthermore, the extent and sophistication of weapons autonomy and its related technologies is likely to develop *incrementally*, thereby enabling commanders to deploy these systems in a wider range of battlefields, against a wider range of targets, and with progressively fewer restrictions on their operation, over time[167]. Conversely, operational systems are unlikely to be developed or fielded in the near-term, due to the likely technological challenges, and the (initial) reluctance of both militaries and ministries of defence to cede control over the use of force at the high (planning) level.

Conclusion on the Sense and Scope of Weapons Autonomy

Three broad conclusions can be drawn from the above. Firstly, as a purely academic concept, 'autonomy' has a number of different meanings, each one being a term of art associated with a particular discipline: politics, philosophy and computer science to name just a few. To retain validity, law and policy discussions concerning LAWS must focus narrowly and squarely on *weapons* autonomy; for example, as defined in this brief note or, at the very least, incorporating some sense of a) humans 'ceding' *tactical* control b) over critical weapons functions c) to complex algorithms that deploying commanders may find relatively opaque; hence, they will need to set *operational* parameters. Arguably, any misapplication of, say, philosophical autonomy to conclude that LAWS development is too remote[168] misses the point of regulating weapons for humanitarian impact, or for strategic stability. Likewise, definitions such as those employed by the UK Ministry of Defence, which set such a high threshold of technical capability as to 'define away' the problem[169], also miss the point of weapons regulation.

Secondly, and following on from this first point, the ultimate aim of defining autonomy in emerging military systems is to delineate a new category of lethal capabilities, which call for new law, or a restatement

---

[164] See 'Metrology: Sensing Sensibility' in *Economist Technology Quarterly* (2017) n. 158, stating that "[q]uantum gravimeters could precisely map geological features from the gravitational force they induce. That would help with getting around in places where satellite-navigation signals are not available – 'a kind of Google Maps for gravitation', as [a British Ministry of Defence scientist] puts it."

[165] See 'Smart Weapons: The Vision Thing', *The Economist* (print edition), 3 December 2016, 67.

[166] 'Raytheon Unveils Compact Multi-Spectral Targeting System', *PR Newswire*, 19 June 2017: http://www.prnewswire.com/news-releases/raytheon-unveils-compact-multi-spectral-targeting-system-300473412.html.

[167] Anderson and Waxman, n. 24; and K. Anderson, D. Reisner, and M. Waxman, M (2014) 'Adapting the Law of Armed Conflict to Autonomous Weapon Systems', *International Law Studies*, Vol. 90, 386.

[168] For example, because machines possess no 'free will' and their actions remain essentially deterministic.

[169] See Ministry of Defence (2011) *The UK Approach to Unmanned Aircraft Systems* (JDN 2/11), page 2-3, describing 'autonomous system' as one that "is capable of *understanding higher level intent* and direction" and that "will, in effect, be *self-aware*", with "*human levels* of situational understanding" (emphasis added): https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/33711/20110505JDN_211_UAS_v2U.pdf.

of the law that *guides commander decision-making* on when to *restrict deployments* or when to *dial up the precautionary measures*. Accordingly, there is a clear distinction between (tactical) autonomy, and (operational and strategic) human control.

Finally, while States are largely not acknowledging any LAWS development programmes, there certainly is a number of tangible indicators pointing to lethal autonomous capabilities in the pipeline. Consequently, it is now important to set an internationally-agreed definition of LAWS, which will delineate the category and assist in creating and / or restating law, as needed.

Comment on The Terminator Dilemma

The remainder of this note will be a brief summary and a few remarks on Paul Scharre's presentation, *The Terminator Dilemma: Autonomous Weapons and the Future of War*, which was delivered at the Barcelona International Workshop on the 'Sense and Scope of Autonomy in Emerging Military and Security Technologies' on 27 February 2017[170].

The LAWS Security Dilemma

Scharre began by establishing that there is currently a strong impetus towards developing and fielding LAWS. This mainly arises from a concern by defence officials that if the United States does not rise to the autonomy challenge, other nations will[171]; and that any asymmetric fielding of LAWS might leave the country at a distinct competitive disadvantage[172]. In this sense, there was an implication that emerging autonomous systems are giving rise to a *security dilemma*; that is, a situation where:

> *"…the self-help attempts of states to look after their security needs tend, regardless of intention, to lead to rising insecurity for others as each interprets its own measures as defensive and measures of others as potentially threatening."*[173]

The core assumptions behind this phenomenon – that global politics is 'anarchic' and, consequently, States must rely on their own means (i.e. 'self-help') – are strongly associated with the *Realist* theory of international relations. The result is a two-level strategic predicament, consisting of a *dilemma of interpretation* and a *dilemma of response*[174]. Moreover, in a climate of mutual mistrust, this manifests itself in an 'offense-defence' conundrum, such that the actual or perceived military build-up of one State (for defensive purposes) may trigger another State (which sees this as offensive) to follow suit. Hence, a *spiral of insecurity* follows in which all enter a destabilising arms race[175]; ultimately, no State is left more secure[176]. Of course, this was a key feature of the Cold War nuclear arms race.

Scharre's recognition of this phenomenon in a LAWS context – also seen in the play on words, *The Terminator Dilemma* – was followed with a crucial question: should the United States (or, indeed, any individual country) go ahead and build these systems? Addressing this question is all the more urgent, given that the basic technologies to build rudimentary LAWS capable of lawful deployment already exists[177].

---

[170] Scharre, n. 30.
[171] Actual comments indicating this approach were cited from Robert Work, *Deputy Secretary of Defense*; and also Frank Kendall, former *Under Secretary of Defense for Acquisition, Technology and Logistics*, US DoD.
[172] Scharre, n. 30. Similar comments were seen above, at (notes and text accompanying) n. 120 and 125.
[173] J.H. Herz, (1951) *Political Realism and Political Idealism*, Chicago: University of Chicago Press, page 7. For some earlier writings which led to this, see J.H.. Herz, (1950) 'Idealist Internationalism and the Security Dilemma', *World Politics*, Vol. 2(2), 157.
[174] That is, how to *interpret* the motives, intentions and capabilities of other States; and how to *respond* to these in the most rational way. See K. Booth, K and N.J. Wheeler, (2008) *The Security Dilemma: Fear, Cooperation and Trust in World Politics*, Basingstoke: Palgrave.
[175] R. Jervis, (1976) *Perception and Misperception in International Politics*, Princeton: PUP, Chapter 3.
[176] Jervis, n.175
[177] Scharre gave the example of autonomous undersea vehicles, which need only be designed to detect large metallic objects for commanders to be confident that the system will engage a military objective (submarine).

However, while Scharre was reluctant to say that weapons autonomy is a positive step forward[178], he fell back to the security dilemma and argued that a unilateral decision by the United States not to develop LAWS will not discourage other countries from pressing ahead; and nor will a pre-emptive ban treaty[179]. Accordingly, the correct question is not '*should* we develop LAWS?' but '*how* to approach the challenge of weapons autonomy, assuming its eventual development and deployment is near-inevitable?'

LAWS Terms and Concepts

From there, Scharre explained the three dimensions of autonomy: the level of human control; machine complexity; and the task being performed by the machine[180]. He also defined and illustrated the three categories of weapons that incorporate autonomy in their critical functions: *fully* autonomous; *semi*-autonomous; and *supervised* autonomous weapon systems[181]. As these terms and concepts have been similarly covered in the above note, there will be no further elaboration here on their descriptive components.

Practical Framework *versus* Thought Experiment

The one aspect that is worth briefly reflecting on is whether the three dimensions might offer a practical framework to delineate LAWS from manned or remotely-piloted weapon systems. It is an important matter because – as has been explained several times throughout this note – weapons autonomy may call for commanders to take stronger and additional precautions in their battle planning; thus, they need to know exactly *which* weapon systems may call for a greater precautionary approach. Scharre did not address this issue directly, but there was a strong sense in his presentation that the third dimension (task being performed) is the only one that offers a tangible indicator of weapons autonomy, which might have policy or planning consequences. Several times, the inadequacy of weapons testing and evaluation (T&E) was offered as a reason not to rely so heavily on these processes, and this may extend to an argument that the second dimension ('machine complexity') should only be considered as a thought experiment or a framework for thinking, but with no empirical reliance on it[182].

Arguably, this would not be a satisfactory place to leave the matter. For as Scharre also argued, 'machine complexity' presents us with a particular paradox: the more intelligent a machine, the greater the range and complexity of tasks that it can perform; but also the harder it becomes to predict the specific actions of that machine in the field, with potentially greater risks for civilians and other protected persons and objects. Accordingly, it is likely that there is a positive correlation between machine complexity and the need for pre-deployment precautions, to effectively mitigate civilian risk. This also chimes with the argument presented above, on more complex autonomous systems exhibiting stochastic (probability-based) reasoning and giving the *appearance* of machine discretion[183]. While not perfect, elaborate T&E processes do exist in order to rate machine complexity[184], with several ways to adapt these to be more relevant to, and valid for, autonomous systems[185]. Furthermore, as was argued above, the subjective element can be addressed (albeit imperfectly) through the legal review process adopting a *reasonable military commander* standard[186]. Thus, there should arguably be no reason to restrict 'machine complexity' to a thought experiment, and if it

---

[178] Arguing that it is "probably a bad thing", in terms of law, ethics, and safety and operational risk.
[179] Towards the end, he also acknowledged the difficulty of gaining consensus towards a pre-emptive ban, and during the Q&A session, Scharre elaborated on the problem of weapons verification, when the principal input for autonomy is intangible software code. Both of these challenges *might* defeat any chance of an enforceable treaty ban on LAWS, as has already been briefly alluded to at (notes and text accompanying) n. 22-25, above.
[180] The account provided above at (notes and text accompanying) n. 39-63 is broadly similar.
[181] Again, the account provided above at (notes and text accompanying) n. 75-79 is broadly similar.
[182] Recall from (note and text accompanying) n. 85, above, that Scharre and Horowitz (and also Crootof) consider the 'automatic', 'automated' and 'autonomous' distinction to be subjective and not administrable.
[183] See (notes and text accompanying) n. 80-83, above.
[184] See (note and text accompanying) n. 86, above.
[185] See (note and text accompanying) n. 90-92, above; though also note the drawbacks contained therein.
[186] See (notes and text accompanying) n. 99-101, above.

is deemed too simplistic to have a binary 'autonomous / non-autonomous' classification, there is nothing to prevent legal review panels from taking a 'levels of machine complexity' approach[187], with as many levels as is deemed necessary to capture the full range of available systems. The aim would be to signal to deploying commanders the level of opaqueness of algorithms and machine behaviour, which was uncovered during the T&E process; hence, the likely restrictions, and the breadth and depth of precautionary measures that may need to be considered during the battle planning process.

Beyond this, Scharre also considered a number of other question, such as: *why* a State might want to build LAWS (security dilemma aside); and what the *legal*, *ethical* and *operational risk* implications might be.

Why Build LAWS?

On the *why* question, he pointed to three main reasons: speed, reliability and mitigating loss of communications. On 'speed', this will clearly be a benefit to each military that can field LAWS, given the super-human data-processing speeds of autonomous systems, and their ability to keep up with the ever-increasing tempo of warfare. An analogy was made with automated stock-trading algorithms, whose split-second reactions have had positive tangible consequences on the bottom line of the firms that employ them. However, an important point that could have been raised is that some part of this 'increased tempo' is, of course, driven by the adoption of increasingly autonomous systems in the first place. Namely, a vicious circle may ensue, whereby LAWS are developed to deal with the increasing tempo; but this results in further rises in tempo, and a self-induced need for yet more increases in weapons autonomy[188].

On 'reliability', Scharre pointed out that one of the reasons that driverless cars are likely to enjoy success – and are currently securing significant policy support from governments around the world – is the removal of human error, and the likely fall in road accidents[189]. In terms of military systems, there is no shortage of recent negative experience that illustrates this same point. For example, there have been numerous incidents involving the wrongful targeting of hospitals[190] or ICRC facilities[191], which were largely down to human error in competences where machines tend to outperform humans. These case examples

---

[187] Similar to the 'levels of autonomy' approach taken by the various reports cited at n. 68-71, above, mostly in the context of the 'human control' dimension.

[188] This 'vicious circle' is arguably at play in the financial markets, with their ever-faster trading algorithms. It was also one of the concerns raised in the 2016 *Meeting of Experts*. See Chairperson's Report, n. 16, ¶ 68.

[189] One estimate put the potential reduction in road traffic accidents at a staggering 90%. See P. Gao, R. Hensley, and A. Zielke, (2014) 'A Road Map to the Future of the Auto Industry', *McKinsey Quarterly*, 2014, Number 4. Available at: http://www.mckinsey.com/insights/manufacturing/a_road_map_to_the_future_for_the_auto_industry.

[190] For example, there was the erroneous US attack on a Médecins Sans Frontières (MSF) clinic in Kunduz, in October 2015, which appeared to directly kill 30 civilians and deny thousands more people critical life-saving treatment. The US commander in Afghanistan (General John Campbell) attributed the targeting mistake mostly to 'avoidable human error'. Among the contributing factors were: pilots not referring to GPS coordinates provided by MSF, which were included on a 'no strike' list; 'fatigue and high operational tempo' endured by troops; and loss of electronic communications on the aircraft. Consequently, the misidentification and erroneous bombing continued despite MSF staff contacting US forces several times *during* the attack to say they were being bombed. Perhaps more telling is the fact that these 'avoidable human errors' were generally in areas where LAWS and machine performance exceeds that of humans'. See 'Kunduz Bombing: US Attacked MSF Clinic 'In Error'', *BBC News*, 25 November 2015: http://www.bbc.co.uk/news/world-asia-34925237.

[191] For example, in the early stages of *Operation Enduring Freedom* in Afghanistan, October 2001, US planes repeatedly bombed ICRC warehouses containing humanitarian supplies, such as food and blankets. Significantly, the warehouses were *within* visual range and "clearly distinguishable from the air by the large Red Cross painted against a white background", but were still attacked by "slow-flying aircraft… [approaching at] low altitude", thereby raising questions as to the situational awareness of the pilots. Again, US authorities were repeatedly notified of the location of the ICRC facilities, the distribution and movement of vehicles, and the gathering of people at distribution points; yet, this did nothing to mitigate the human targeting errors, with the result that 55,000 disabled civilians were denied vital resources for food and warmth. See ICRC *Press Release 01/43*, 16 October 2001: https://www.icrc.org/eng/resources/documents/news-release/2009-and-earlier/57jrcz.htm; and ICRC *Press Release 01/48*, 26 October 2001: https://www.icrc.org/eng/resources/documents/news-release/2009-and-earlier/57jrdx.htm.

arguably provide a strong normative basis for the fielding and judicious deployment of LAWS, especially as many of those human errors tend to be simple omissions, like failing to check the crucial 'no strike list' before weapons release: an easy oversight to make by war-weary humans; yet, a highly routine task with near-perfect compliance by an autonomous system.

Finally, Scharre elaborated on 'loss of communications', which are most likely to occur in symmetric (high-intensity) environments, where enemy jamming is rife. He raised some important questions that weapons developers and deploying commanders will need to consider in advance. Namely, what should a LAWS be programmed to do in the event of lost communications? Should it return to base? Continue with non-lethal activities, such as reconnaissance? Switch to semi-autonomous mode, and strike pre-approved targets only? Or should it continue to operate autonomously, actively searching for targets over a wide area? Given the clearly contextual nature of the matter, it is not possible to fix an answer in advance. Instead, the answer is likely to depend on: the precise capabilities of the system being deployed, and its reliability; the nature of the battlefield; and the extent of military advantage to be gained from continued operation, which will determine the extent of civilian risk that may be lawful.

One important point that was not raised on the question of *why* States will build LAWS is the cost factor. However, this is crucial, especially given the recent experiences of austerity in government spending throughout the West. These have led to defence budget cuts, for example, in the American *Budget Control Act 2011*[192], which set defence sequestration caps. Furthermore, while the Trump Administration intends to boost US military spending[193], it may experience problems achieving this in the short-run[194]; and, in any event, long-run cost-efficiency remains a concern, as was seen in the most recent analysis of the US Administration's five-year defence plan[195]. These and other cost considerations will very likely be a driving force for LAWS adoption, both in the US and in other NATO countries. But also elsewhere, such as in Russia, where structural weaknesses in the economy and a continued confidence crisis have resulted in seven recent quarters of economic contraction, and have undermined growth in the defence budget[196].

Legal Considerations

On legal matters, Scharre pointed out that there is nothing in IHL that mandates manual targeting over autonomous targeting. The law only concerns itself with actual or anticipated *effects* on the battlefield (via the principles of distinction and proportionality); and pre-emptive or corrective *actions* (like precautionary measures, both before and during deployment). Accordingly, he emphasised that a) if a LAWS can be deployed and used in a manner that meets these criteria, then the weapon system will be *a priori* lawful; and b) the only asymmetry is that LAWS are not legal persons, thus cannot be 'bound' by IHL, as such. Only humans are the addressees of the law of war, so human commanders and operators will always have to make the judgment call on whether a robot's actions are lawful. Both of these points have already been made by, for example, Schmitt[197], Thurnher[198] and Anderson et al.[199], amongst others. Yet, the first is a not a universally held position, and with ongoing discussion of a pre-emptive ban, it remains an important point to reiterate. Scharre's corollary is that these will have important implications for the design and use of LAWS,

[192] See M. Moyar, 'How Obama Shrank the Military', *The Wall Street Times*, 2 August 2015: https://www.wsj.com/articles/how-obama-shrank-the-military-1438551147.
[193] P. Scharre, and A. Routh, 'President Trump's Defense Spending Request', *CNAS Press Note*, 28 February 2017: https://www.cnas.org/press/press-note/cnas-press-note-president-trumps-defense-spending-request.
[194] A. Lowrey, 'Why Sequestration is Poised to Kill Trump's Budget', *The Atlantic*, 16 March 2017: https://www.theatlantic.com/politics/archive/2017/03/donald-trump-meet-sequestration/519798/.
[195] Congressional Budget Office, *An Analysis of the Obama Administration's Final Future Years Defense Program*, April 2017, projecting 2017 FYDP costs through to 2032: https://www.cbo.gov/system/files/115th-congress-2017-2018/reports/52450-fydp.pdf.
[196] S. Oxenstierna, (2016) 'Russia's Defense Spending and the Economic Decline', *Journal of Eurasian Studies*, Vol. 7(1), 60.
[197] Schmitt, . 27.
[198] Thurnher, n. 115.
[199] Anderson et al. n. 167.

and this can certainly be seen in Sharkey's model, for instance, which proposes a five-level taxonomy of 'human-machine collaboration', with the explicit aim of designing systems with greater humanitarian impact[200].

Ethical Considerations

Scharre considered ethical concerns to be more problematic, and relevant to LAWS on two levels. Firstly, what is 'legal' and what is 'right' is not always one and the same thing. This issue has been raised by many LAWS critics, beginning with the Human Rights Watch report, *Losing Humanity*[201], which raised concerns about robots engaging in status-based targeting in situations where a human soldier may have spared the enemy combatant. For example, where the latter has neither surrendered nor been incapacitated, but is unlikely to engage in combat and could reasonably be captured. Lethal targeting in such a case is both *legal* and *likely* to be carried out by a LAWS programmed with objective parameters; but the more complex reasoning of the human combatant, which may invoke empathy and consider ethical imperatives, may well lead to a voluntary decision to refrain from using lethal force.

Secondly, even some *legal* decisions require value judgments that do not have a clear answer, but which must be addressed with an appeal to ethical reasoning. Scharre gave the example of proportionality assessments, which must ensure that an attack does not involve estimated collateral damage and incidental injury to civilians[202] that is 'excessive' in relation to the concrete and direct military advantage anticipated[203]. While collateral damage estimation is a highly quantitative process, which can surely be automated and integrated into a LAWS[204], the assessment of military advantage is both subjective and contextual[205]; and the meaning of 'excessive' is even more contextual and is inherently indeterminate[206]. Consequently, many would argue that a human must assess the proportionality of an attack. Concretely, Scharre raised the question of whether military professionals can ever offload such decisions to a machine, which at best will invoke *ex ante* human decisions made by a software programmer, in an abstract setting, far removed from the current battlefield.

As before, however, the answer is likely to be found in 'human-machine collaboration' that is concerned with system design and task allocation. Such efforts aim to design LAWS in a way as to combine deliberative human reasoning with the speed and efficiency of autonomous data-processing. In the case of proportionality assessments, a remote 'dial-in' capacity might enable commanders to update systems with changing values for the military advantage anticipated (MAA), while each LAWS assesses its own collateral damage estimate (CDE).

In this regard, Van den Boogaard (2015) suggests a highly sophisticated solution, based precisely on having such a 'dial-in' capacity[207]. The underlying approach of the author is to focus on the *level of command*[208] at which proportionality assessments should take place. Accordingly, he envisages a scenario

---

[200] See Sharkey, n. 71, whose model aims not to promote LAWS, but to retain meaningful human control.
[201] Human Rights Watch and the IHRL Clinic, Harvard Law School, n. 15.
[202] Note that collateral damage (to protected objects) is different to incidental injury (to civilians and other protected persons), but the following will refer to 'collateral damage' as shorthand to refer to both.
[203] Articles 51(5)(b) and 57(2)(a)(iii), *AP I*.
[204] Schmitt, n. 27.
[205] Sassóli (n.28) points out that the military advantage of a given attack may be accurately known on deployment, but then shifts rapidly depending on the plans of the commander and the development of military operations on both sides; and this poses the most serious IHL argument against full and prolonged autonomy.
[206] "By American domestic law standards…[IHL] proportionality…would be constitutionally void for vagueness": W.Hays Parks, (1990) 'Air War and the Law of War', *Air Force Law Review*, Vol. 32(1), 1, page 173.
[207] See J. Van den Boogaard, (2015) 'Proportionality and Autonomous Weapons Systems', *Journal of International Humanitarian Legal Studies*, Vol. 6(2), 247, especially pages 275-277.
[208] As explained in the *DoD Dictionary of Military and Associated Terms*, March 2017, there are three levels:

with strong communication links and elaborate data flows: both *horizontally*, between tactical autonomous units; and *vertically*, between those same units and the operational and strategic headquarters[209]. In such a case, it is possible to continuously update the situation on each level, thereby deriving time-sensitive tactical, operational and strategic proportionality assessments. In particular, multilateral data flows and feedback enable each level of command to update the assessments of CDE, MAA and, therefore, 'proportionality' – all in *real time*. For example, if an attack by a tactical autonomous unit derives a military advantage and gets closer to achieving the goals of the larger operation, this will decrease the assigned value of the MAA that other parts of that same operation are intended to achieve, thereby reducing the likelihood that a disproportionate attack may be planned at the operational level[210]. Likewise, if changes in strategic priorities occur at the political level, these can be manually inputted at strategic headquarters, thereby updating the MAA of targets assigned to each tactical unit; importantly, the human-inputted MAA data can be 'converted' into a CDE threshold, for autonomous operation. Accordingly, human commanders remain in control at the broader *operational* and *strategic* levels, due to the need for highly qualitative assessments involving strategic and political factors at those levels[211]. Concurrently, rapid and data-heavy proportionality calculations are autonomously done at the *tactical* level, which can sometimes move too fast for human operators to keep pace[212]. This enables (tactical) 'narrow loop' autonomy to proceed alongside deliberative human decision-making in the 'wider loop' of (strategic and operational) control[213].

This approach has both intuitive appeal and implicit support from other commentators[214], yet it poses at least one major limitation: the system would require reliable and continuous communication links, which may be possible in an asymmetric battlefield where the LAWS deploying side has massive technological advantage; but it cannot always be guaranteed. Indeed, as Scharre pointed out earlier, one of the main reasons for adopting LAWS is precisely to operate in denied environments in a symmetric (high-intensity) battle, where such links are either not available, or where they would expose systems to an unacceptable risk of enemy hacking[215]. That said, with the recent reassignment of cyber security to the top of the Pentagon's agenda[216], there may be less reason in future to doubt the robustness of communication links[217]. Thus, with an appropriate system design that optimises task allocation between man and machine – and assuming communications superiority – there remain some possibilities for designing-in human discretion in areas that call for ethical considerations and deliberative decision-making.

- **Strategic** level: where higher *national* or *multinational* security objectives and guidance are determined, then *national resources* are developed and used to achieve those objectives.
- **Operational** level: where *campaigns* and major operations are planned, conducted and sustained to achieve strategic objectives within *specific theatres* or other *operational areas*.
- **Tactical** level: where *individual battles and engagements* are planned and executed to achieve military objectives assigned to *tactical units* or *task forces*.

[209] Van den Boogaard, n. 207.
[210] Van den Boogaard, n. 207.
[211] Van den Boogaard, n. 207.
[212] Again, note that autonomous proportionality calculations here are derived from: *human-inputted* MAA data; conversion of that data to a CDE threshold; and actual CDE assessments done by onboard sensors and software.
[213] See (notes and text accompanying) n. 102-114, above, on 'wide' and 'narrow' loops.
[214] See, for example, See Kalmanovitz, P 'Judgment, Liability and the Risks of Riskless Warfare', in Bhuta et al. (2016) n. 71, 145-163, discussing the linkages between tactical and strategic military advantage; how LAWS may be able to estimate the former; and why the latter implicates political goals, thereby necessitating some human involvement in proportionality calculations.
[215] Other potential problems include bandwidth limitations, which may provide insufficient capacity for such data-heavy transmissions; and simple communication breakdowns from potentially unreliable or in-battle damaged components. These are not necessarily persistent problems, but they do point to the need for rigorous testing of both communications hardware and software, to be assured of reliable battlefield performance.
[216] J. Keller, 'Iran-US RQ-170 Incident Has Defense Industry Saying 'Never Again' to Unmanned Vehicle Hacking', *Military & Aerospace Electronics*, 3 May 2016, noting that the Pentagon is exploring techniques like multi-level encryption, to try to make unmanned systems as close to being hack-resistant as possible: http://www.militaryaerospace.com/articles/2016/05/unmanned-cyber-warfare.html.
[217] Though this is currently a mere hope, and does not foresee the full range of possible enemy counter-measures in the future, which may neutralise multi-level encryption and reintroduce the risk of cyber-attacks on LAWS.

Operational Risk

Finally, Scharre raised the question of how much trust we should put into autonomous systems and, by extension, how much trust to put into testing and evaluation (T&E) processes. T&E are never 100% failsafe, yet by delegating more legal authority to a LAWS, we implicitly place more trust in these imperfect processes. Should there be a malfunction or an unexpected system interaction that is not detected during T&E, a LAWS can ultimately do more harm than good. This was vividly illustrated in the 2003 *Patriot* fratricide incidents[218], which could conceivably occur again on a grander scale and involve civilian casualties. Moreover, should such incidents occur in an *adversarial* context, this may lead to even more serious consequences that undermine strategic stability; such as the hypothetical scenario of unanticipated interactions between LAWS deployed by different adversaries, which rapidly escalates into a 'flash war'[219].

Scharre ended by saying that States must do two things to minimise operational risk and maintain strategic stability.

- Firstly, they should *individually* implement failsafe measures, similar to the circuit-breakers that prevent stock-trading algorithms from damaging interactions. Such mechanisms are often triggered by pre-determined actions and patterns that are judged to signal the potential onset of a flash crash; or, in this case, a flash war.
- Secondly, States should *collectively* agree on best practices and 'rules of the road' on how autonomous systems should interact. This is to avoid misperceptions between LAWS, again to prevent unwarranted engagements that might escalate into a flash war.

Scharre did not comment on what shape these rules or best practices might take, and indeed it would have been beyond the scope of his presentation to look into legal transplants or the like. However, it is worth briefly mentioning that the current framework of arms control agreements provides useful examples of States agreeing to mutually beneficial 'rules of the road' that promote strategic stability. Many of these rules have been applied to conventional and nuclear arms alike, and they can be grouped into three categories.

- Quantitative limits on treaty-accountable items, like military hardware and armaments.
- Spatial restrictions (regional, geographic and zonal).
- Functional measures, mostly aimed at confidence- and security-building[220].

Briefly, by agreeing on *quantitative limits* on treaty-accountable LAWS, States would primarily aim to address the security dilemma and to allay fears of a destabilising arms race, with secondary effects on operational risk. This particular type of rule was prominent in the *Treaty on Conventional Armed Forces in Europe*[221] (*CFE Treaty*), which put upper limits on five categories of 'Treaty Limited Equipment' (TLE) between NATO and the former Warsaw Treaty Organisation[222]. It also imposed further caps on 'active deployment'[223], to limit more directly the extent of combat readiness. Quantitative limits are also a key feature of many

[218] See (note and text accompanying) n. 59
[219] Scharre has written about this in greater depth in Scharre, n. 9.
[220] P.R. Viotti, 'A Template for Understanding Arms Control' in R.E. Williams Jr, and P.R. Viotti, (eds.) (2012) *Arms Control: History, Theory, and Policy*, Oxford: Praeger Security International, 7-14.
[221] *Treaty on Conventional Armed Forces in Europe*, 19 November 1990, 30 ILM 1, entered into force 17 July 1992.
[222] Article IV(1), *CFE Treaty*, established five TLE categories with numerical 'ceilings' allotted to each Group of States to 20,000 battle tanks; 30,000 armoured combat vehicles; 20,000 pieces of artillery; 6,800 combat aircraft; and 2,000 attack helicopters.
[223] In line with Article IV(1), *CFE Treaty*, which put further limits on 'active deployment' within the first three TLE categories of battle tanks, armoured combat vehicles and artillery pieces.

bilateral nuclear arms control treaties, such as *New START*[224], which mandated reductions and limits between the US and Russia in three nuclear-related categories[225]. As mentioned above, this particular kind of rule primarily addresses the security dilemma; however, it could also be argued that by collectively limiting both aggregate and deployed LAWS, States will ease the burden of monitoring interactions between algorithms and, therefore, better manage the operational risk.

Spatial restrictions were also a feature of the *CFE Treaty*, which imposed *regional limitations* on deployment by way of three *concentric zones*[226] and a *flank zone*[227], which made it difficult to build up forces near the former 'Iron Curtain' or its flanks. The aim of these arrangements was also to pre-empt an arms race or any destabilising concentrations of conventional arms, thereby alleviating fears of a surprise attack[228]. In a LAWS context, a prohibition on loitering within a certain range of an adversary's airspace may allay suspicions of 'aerial occupation', and maintain a healthy distance between autonomous systems that may be prone to misperceiving each other's actions. Nuclear treaties that take a similar approach include the *Antarctic Treaty*[229], the *Outer Space Treaty*[230] and the *Seabed Arms Control Treaty*[231]. Each one prohibits the emplacement of nuclear weapons in the specified environment, largely to prevent international armed conflict and nuclear tensions creeping into areas that that were hitherto free from these. While conceptually different, this is similar to why States might want to restrict the geographical deployment of adversaries' LAWS[232]: to prevent areas immediately outside their airspace and territorial waters – namely, areas currently difficult or costly to access for extended periods by manned weapon systems – from becoming permanently 'occupied' and constantly the subject of monitoring and analysis by their own autonomous systems.

Finally, *functional measures* may offer the most compelling examples of commonly-agreed rules that aim to manage the operational risk associated with LAWS deployments. These include providing advance notification of military exercises to obviate any expectation of an impending attack; an example being the *Helsinki Final Act*[233], which mandated prior notification of major and other military manoeuvres between 35 State Parties during the height of the Cold War. Similarly, the *Ballistic Missile Launch Notification Agreement*[234] required each of its Parties (the US and the Soviet Union) to provide each other no less than 24 hours' notice of the planned date, launch area and area of impact for any launch of a strategic ballistic missile[235]. Again, the aim was to mitigate the risk of (nuclear) war as a result of misinterpretation, miscalculation, or an accident. Applied in a LAWS context, the notification of training exercises and unusual deployments of autonomous systems should remove the unwanted element of surprise and enable adversaries to perceive innocuous deployments for what they really are; in turn, the algorithms of their own

---

[224] *Treaty Between the United States of America and the Russian Federation on Measures for the Further Reduction and Limitation of Strategic Offensive Arms*, US-Russia, 8 April 2010, S. Treaty Doc. No. 111-5, entered into force 5 February 2011 (*New START*).
[225] Article II, *New START*, limits aggregate numbers to: 800 deployed and non-deployed nuclear launchers and heavy bombers; within that limit, 700 deployed missiles and heavy bombers; and 1550 deployed warheads.
[226] Article IV(2)-(4), *CFE Treaty*, which put the tightest restrictions at the 'centre' of Europe and progressively loosened the quantitative ceilings for each successive zone moving outwards.
[227] Article V(1), *CFE Treaty*, which put separate quantitative ceilings in the northern and southern flank areas.
[228] Preamble, *CFE Treaty*, which stated its aims as "…eliminating disparities prejudicial to stability and security and…eliminating the capability for launching surprise attack and for initiating large-scale offensive action in Europe."
[229] *Antarctic Treaty*, 1 December 1959, 12 UST 794, entered into force 23 June 1963.
[230] *Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies*, 27 January 1967, 18 UST 2410, entered into force 10 October 1967.
[231] *Treaty on the Prohibition of the Emplacement of Nuclear Weapons and Other Weapons of Mass Destruction on the Seabed and the Ocean Floor and in the Subsoil Thereof*, 11 February 1971, 23 UST 701, entered into force 18 May 1972.
[232] The conceptual difference arises from the fact that the three treaties address areas that are more difficult to reach than the outer perimeter of an adversary; and, with Antarctica and the seabed, less militarily relevant too. This is likely to make it more politically difficult to transplant these rules into a LAWS treaty.
[233] Organization for Security and Co-operation in Europe (OSCE), *Conference on Security and Co-operation in Europe: Final Act of Helsinki*, 1 August 1975.
[234] *Agreement Between the United States of America and the Union of Soviet Socialist Republics on Notifications of Launches of Intercontinental Ballistic Missiles and Submarine-Launched Ballistic Missiles*, 31 May 1988, 27 ILM 1200, entered into force 31 May 1988.
[235] Article I, *Ballistic Missile Launch Notification Agreement*.

LAWS may be updated accordingly, or put on 'man-on-the-loop' mode, until the unusual deployments of the adversary have passed.

Subsequent to the *Notification Agreement*, *START I*[236] entered into force and expanded its notification requirements to include additional telemetry-related information from missiles[237]. While *START I* is no longer in force, this aspect is retained in Article IX of *New START*; furthermore, Paragraph 5(h) of its *Annex on Telemetric Information* provides for the exchange of encoding algorithms, to assist in decoding and analysing all telemetric data provided.

In a LAWS context, the sharing of *certain* vital information – such as whether a LAWS deployment is armed or unarmed, for surveillance purposes – and the sharing of *certain aspects* of LAWS algorithms[238] between potential adversaries may enable each one to properly understand the capabilities of the other side; and for each of their autonomous systems to properly distinguish offensive behaviour from genuinely non-threatening behaviour[239]. This may move systems closer to 'peacetime compatibility', and go some way towards mitigating the risk of accidental or 'flash' war as a result of automated misinterpretations. Such agreements are arguably more likely to be reached where there are already improving diplomatic relations between potential adversaries[240]; as was the case, for example, with NATO and the Warsaw Treaty Organisation when the *CFE Treaty* was signed. Yet, they are also never guaranteed to work, as there may remain a number of residual risks even after any such agreements have been reached[241].

However, while the above rules and agreements provide ample evidence of what has been achieved in the past, these were mostly in nuclear and Cold War contexts; it may be doubtful that States will be so quick to develop similar rules in a LAWS context. Indeed, as one scholar commented, LAWS do not (currently) evoke the same visceral responses as do images of the mushroom cloud, which were so instrumental to the success of nuclear arms control[242]. Thus, unless States take this issue more seriously, pre-emptive action in the short-run to develop 'rules of the road' is likely to be a long-shot; in which case, it may well be an actual flash war that motivates them to take seriously the idea of LAWS arms control.

---

[236] *Treaty on the Reduction and Limitation of Strategic Offensive Arms*, US-USSR, 31 July 1991, S. Treaty Doc. No. 102-20 (1991), entered into force 5 December 1994 (*START I*).

[237] Articles VIII(3)(f) and X, *START I*; *Protocol on Telemetric Information*. These include broadcast frequencies, modulation types, and whether encapsulation or encryption would be used during flight tests.

[238] There is no suggestion that classified and security-sensitive aspects of LAWS algorithms can, or should be shared. Instead, the focus is on where States believe that a) their systems have some innocuous algorithmic features; which b) may nonetheless be misperceived by rival LAWS; and c) these algorithmic features are both distinct to, and separable from, the classified aspects of their systems. Where these conditions are *cumulatively* met, it may be in the mutual self-interest of States to share such details, so that rival algorithms move closer towards 'peacetime compatibility', thereby mitigating the risk of a flash war that no party wants.

[239] For example, consider a swarm of drones flying in formation at a particular speed and altitude, purely as a training exercise or for surveillance purposes. In the case of the latter, swarm surveillance may be objectively justified to obtain panoramic images, with data from all drones being combined to derive the clearest possible picture. Yet, the adversary's own autonomous systems may interpret the particular combination of swarming, airspeed and altitude as an offensive manoeuvre – possibly even an impending attack – which calls for an armed response in putative self-defence. By sharing these innocuous aspects of their LAWS algorithms, each State will assist the other in updating the latter's own algorithms, to avoid potential mishaps.

[240] M. Trachtenberg, (1991) 'The Past and Future of Arms Control', *Daedalus*, Vol. 120(1), 203, page 212, describing arms control agreements more generally as "icing on the cake of political accommodation", which bolsters the peace by acting as a reward for, and reinforcement of, friendly political behaviour.

[241] For example, there may remain a number of unforeseen situations which, almost by definition, cannot be planned for; if so, these may still result in unanticipated interactions, leading to a flash war. Also, depending on the political and diplomatic context, it is possible that receiving States would still assume that an adversary may continue preparing ruses, such as a surprise attack designed to appear as one of the 'innocuous manoeuvres'; if so, the receiving State may ignore the transmitted data and prefer to retain its autonomous capacity for an armed response in such a scenario.

[242] N.K. Modirzadeh, (2014) 'Autonomous Weaponry and Armed Conflict', *Panel Discussion of the American Society of Int'l Law*, 10 April 2014. Full video available at: https://www.youtube.com/watch?v=duq3DtFJtWg.

Furthermore, with NATO-Russia relations becoming increasingly tense since the annexation of Crimea in 2014[243], it may be difficult to imagine certain adversaries negotiating rules of the road on LAWS any time soon. Such tensions include: NATO's practical suspension of cooperation with Russia[244]; Russia's *effective* withdrawal from the *CFE Treaty* in March 2015, partly as confirmation of its tensions with NATO and a desire to close down channels of communication[245]; a serious impasse between the US and Russia over alleged violations of the *Intermediate-Range Nuclear Forces Treaty*[246], which has led to accusations being levelled to and from both sides[247]; and more recent withdrawals by Russia from certain pacts and agreements, both nuclear-related[248] and conventional[249]. The latter contained a hint of irony, as the Memorandum of Understanding that was suspended by Russia created certain protocols in order to "minimiz[e] the risk of inflight incidents" in crowded Syrian airspace[250]; as such, it aimed to mitigate similar operational risks as those posed by LAWS. Accordingly, the political climate is not currently conducive to international cooperation on such issues as sharing certain details of algorithms, or broader rules of the road on weapon systems. Thus, it is fair to expect that cooperation on LAWS arms control may remain rather muted, at least in the short-run.

As a final point of caution, there is a number of more deep-rooted or 'structural' reasons why a standalone LAWS arms control treaty may be unlikely. It is beyond the scope of this comment to go into these, but suffice it to say that so long as LAWS will not be able to carry out major and sustained offensive operations that will replace combined arms operations, a standalone arms control agreement on LAWS may be unrealistic[251]. That said, some of the above provisions may well be the subject of an additional protocol to an existing arms control agreement. Alternatively, a simpler clarifying statement may subsume LAWS capabilities into existing instruments[252].

Conclusion on The Terminator Dilemma

---

[243] See NATO 'NATO-Russia Relations: The Background', *NATO Media Backgrounder*, February 2017: http://nato.int/nato_static_fl2014/assets/pdf/pdf_2017_02/20170206_1702-nato-russia-en.pdf.

[244] 'Ukraine Crisis: NATO Suspends Russia Cooperation', *BBC News*, 2 April 2014: http://www.bbc.co.uk/news/world-europe-26838894.

[245] K. Hille, and N. Buckley, 'Russia Quits Arms Pact as Estrangement with NATO Grows', *Financial Times*, 10 March 2015: https://www.ft.com/content/f6c814a6-c750-11e4-9e34-00144feab7de. It is worth noting that this is not the first time that Russia has thrown the future of the *CFE Treaty* into doubt, having also suspended their membership in 2007 for a variety of NATO actions back then.

[246] *Treaty Between the United States of America and the Union of the Soviet Socialist Republics on the Elimination of Their Intermediate-Range and Shorter-Range Missiles*, US-USSR, 8 December 1987, 1657 UNTS 485, entered into force 1 June 1988.

[247] U. Kühn, and A. Péczeli, (2017) 'Russia, NATO, and the INF Treaty', *Strategic Studies Quarterly*, Vol. 11(1), 66, citing US accusations that Moscow has tested ground-launched cruise missiles (GLCMs) within the ranges banned by the *INF Treaty*, and concerns that Russia is producing more missiles than is needed to sustain a flight-test programme. For its part, Russia has cited extensive US use of armed drones, which it considers to be a GLCM; and deployment in Poland and Romania of the MK 41 Vertical Launch System, which Russia considers can be used to launch intermediate-range cruise missiles.

[248] Namely, the *2000 Plutonium Management and Disposition Agreement* as amended by the *2010 Protocol*. See A.E. Kramer,. 'Vladimir Putin Exits Nuclear Security Pact, Citing 'Hostile Actions' by the US', *The New York Times*, 3 October 2016, also citing the alleged inability of the US to dispose of excessive weapons plutonium: https://www.nytimes.com/2016/10/04/world/europe/russia-plutonium-nuclear-treaty.html?_r=0.

[249] M. Eckstein, 'Russia Suspends Air Space Deconfliction Agreement with US after Chemical Weapons Retaliation Strikes', *USNI News*, 7 April 2017: https://news.usni.org/2017/04/07/russia-suspends-air-space-deconfliction-agreement-with-u-s-navy-osd-pushing-for-continued-safety-related-dialogue.

[250] L. Ferdinando, 'US, Russia Sign Memorandum on Air Safety in Syria', *DoD News*, 20 October 2015: https://www.defense.gov/News/Article/Article/624964/us-russia-sign-memorandum-on-air-safety-in-syria/.

[251] Indeed, a historical look at the *CFE Treaty* reveals that State Parties were convinced that major offensives at the time could only be carried out via combined arms operations involving the five TLE categories (see n. 222); hence, the impetus to negotiate such a comprehensive agreement. With nuclear arms control, the catastrophic risk associated with a single attack provided enough impetus for States to negotiate dedicated nuclear agreements. This historical fact illustrates that discrete arms control agreements tend to focus on discrete or combined offensive capabilities that have a sufficiently grave and sustained impact. It is doubtful that LAWS, as a single category of military capability, will be at that stage for some time to come.

[252] I am grateful to Wolfgang Richter for pointing out these practical limitations and their potential solutions.

Overall, Scharre's presentation offered an excellent introduction to both the technical and definitional aspects of autonomy in weapon systems, and it raised a number of important questions on the normative issues that States will need to consider as we move closer towards these capabilities. The most compelling aspect of the presentation was on the operational risk factors, which will certainly raise concerns for strategic stability. His call for States to develop 'rules of the road' is both timely and appropriate, given: the apparent sense of military competition in LAWS development; the likely complexity of autonomous systems, especially in an adversarial context; and the algorithmic 'crashes' already seen in the commercial sector. That said, given the current diplomatic situation between the US / NATO and Russia – not to mention the many political tensions between other potential adversaries – there are likely to be some challenges in developing rules of the road, at least until the current set of impasses are addressed to a satisfactory level. Even then, more 'structural' conditions may mean that a comprehensive LAWS arms control agreement may be unlikely, thus amendments, additional protocols or clarifications to existing instruments may be the most likely solution in today's political climate.