The Rio Seminar on Autonomous Weapons Systems, held in Rio de Janeiro at the Naval War College on February 20, 2020, aimed at contributing to the debate on the governance of emerging technologies in LAWS (Lethal Autonomous Weapons Systems) under international law, including IHL (International Humanitarian Law).

The Rio Seminar took place in the framework of the GGE-LAWS of the CCW (Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons).

Its purpose was to foster discussions among the main participants of the LAWS negotiations—government representatives, international organizations, International Committee of the Red Cross, non-governmental organizations, private sector, and academia—in a multi-stakeholder approach considering its diplomatic, legal, technological, corporate, strategic, and military dimensions. The informal setting enabled a dynamic knowledge sharing, which may help governments and non-governmental delegations in preparing for the GGE activities in 2020, and its recommendations to the next Meeting of the High Contracting Parties, in 2020, and the Sixth Review Conference of the CCW, in 2021.

The video presentations of the Rio Seminar are available at: <https://m.youtube.com/playlist?list=PLY4MsNDouGfge7-IAdRZtdJk2mJrwljVz>.

FUNDAÇÃO ALEXANDRE DE GUSMÃO
www.funag.gov.br

ISBN 978-65-87083-30-8
9 786587 083308

Rio Seminar on Autonomous Weapons Systems

FUNDAÇÃO ALEXANDRE DE GUSMÃO

coleção
EVENTOS

Rio Seminar on Autonomous Weapons Systems

Rio de Janeiro • Naval War College
February 20th, 2020

Fundação Alexandre de Gusmão

coleção
EVENTOS

Rio Seminar on Autonomous Weapons Systems

Rio de Janeiro • Naval War College
February 20th, 2020

Fundação Alexandre de Gusmão

coleção
EVENTOS

Rio Seminar on Autonomous Weapons Systems

Ministry of Foreign Affairs
Alexandre de Gusmão Foundation

The Alexandre de Gusmão Foundation – FUNAG, established in 1971, is a public foundation linked to the Ministry of Foreign Affairs whose goal is to provide civil society with information concerning the international scenario and aspects of the Brazilian diplomatic agenda. The Foundation's mission is to foster awareness of the domestic public opinion with regard to international relations issues and Brazilian foreign policy.

FUNAG is headquartered in Brasília, Federal District, and has two units in its structure: the International Relations Research Institute – IPRI, and the Center for History and Diplomatic Documentation – CHDD, the latter being located in Rio de Janeiro.

Rio Seminar on Autonomous Weapons Systems

Brasilia – 2020

Ministry of Foreign Affairs
Esplanada dos Ministérios, Bloco H, Anexo II, Térreo
70170-900 Brasília-DF
Phone numbers: +55 (61) 2030-9117/9128
Website: www.funag.gov.br
E-mail: funag@funag.gov.br

Editorial staff:
Acauã Lucas Leotta
Diego Marques Morlim Pereira
Gabriela Del Rio de Rezende
Luiz Antônio Gusmão

Review:
Guilherme Lucas Rodrigues Monteiro

Graphic design and cover:
Varnei Rodrigues - Propagare Comercial Ltda.

Dados Internacionais de Catalogação na Publicação (CIP)

R585 Rio Seminar on Autonomous Weapons Systems (2020 February 20 : Rio de Janeiro, Naval War College) - Brasília : FUNAG, 2020.

321 p. – (Coleção Eventos)

Seminário sobre Sistemas de Armas Autônomas realizado no Rio de Janeiro, na Escola de Guerra Naval em 20 de fevereiro de 2020.

ISBN 978-65-87083-30-8

1. Sistemas autônomos – armas - robôs. 2. Inteligência artificial. 3. Direito internacional. I. Título.

CDU 681.5
CDD 629.89

Depósito legal na Fundação Biblioteca Nacional conforme Lei no 10.994, de 14/12/2004.
Bibliotecária responsável: Raimunda Lima Evangelista, CRB-1/3382

Contents

Panel 1
Human-Machine Interaction and Human Control: From Engineering to IHL

Panel 2
International Law, Including IHL, on Laws: Is There a Need for a New Protocol?

List of Abbreviations

ABIMDE	Brazilian Defense and Security Industries Association
AI	Artificial intelligence
AP-I	Additional Protocol I (to the 1949 Geneva Conventions)
AWS	Autonomous weapon systems
BND	Germany's Federal Intelligence Service
C4ISR	Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance
CAII	Consumer Artificial Intelligence Information Leaflet
CCW	Convention on Certain Conventional Weapons
CEIA	Certificação em Ética para Inteligência Artificial (Certification in Ethics for Artificial Intelligence)
CIA	United States Central Intelligence Agency
C-RAM	Counter-Rocket Artillery and Mortar system

DARPA	Defense Advanced Research Projects Agency
DIB	Defense Industrial Base
DIPROD/MRE	Division for Products on Defense – Brazilian Ministry of Foreign Affairs
DoD	Department of Defense (United States)
EC	European Commission
EGN	Brazilian Naval War College
ESG	Environmental, social and governance
ESI	Emergency Severity Index
FUNAG	Alexandre de Gusmão Foundation
GGE-LAWS	Group of Governmental Experts on Lethal Autonomous Weapons Systems
HC	Human control
HITL	Human in the Loop
HMI	Human-machine interaction
HOOTL	Human out of the Loop
HOTL	Human on the Loop
ICRC	International Committee of the Red Cross
IEEE	Institute of Electrical and Electronics Engineers
IHL	International Humanitarian Law
IHRC	International Human Rights Clinic
IL	International Law

iPRAW	International Panel on the Regulation of Autonomous Weapons
ISR	Intelligence, Surveillance, and Reconnaissance
LAWS	Lethal Autonomous Weapons Systems
NPT	Treaty on the Non-Proliferation of Nuclear Weapons
NSA	United States National Security Agency
PIL	Patient Information Leaflet
SEIPRODE	Special Inter-Ministerial Secretariat
SHIELD	Supply Chain Hardware Integrity for Electronics Defense program
SIPRI	Stockholm International Peace Research Institute
TPNW	Treaty on the Prohibition of Nuclear Weapons
U.S.	United States
XAI	Explainable AI

Rio Seminar on Autonomous Weapons Systems

Rio de Janeiro, Naval War College
February 20, 2020

Presentation

The Rio Seminar on Autonomous Weapons Systems, held in Rio de Janeiro at the Naval War Collegeon February 20, 2020, aimed at contributing to the debate on the governance of emerging technologies in LAWS (Lethal Autonomous Weapons Systems) under international law, including IHL (International Humanitarian Law).

The video presentations of the Rio Seminar are available at: <http://www.funag.gov.br/index.php/en/news/3072- egistrationsopen-for-the-rio-seminar-on-autonomous-weapons-systems>.

I wish to express my gratitude to the Naval War College (Brazilian Navy) and the Alexandre de Gusmão Foundation (FUNAG, Ministry of Foreign Affairs of Brazil) for the extraordinary support in the organisation and conduction of the whole event.

The Rio Seminar took place in the framework of the GGE-LAWS of the CCW (Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons).

Its purpose was to foster discussions among the main participants of the LAWS negotiations—government representatives, international organizations, International Committee of the Red Cross, non-governmental organizations, private sector, and academia—in a multi-stakeholder approach considering its diplomatic, legal, technological, corporate, strategic, and military

dimensions. The informal setting enabled a dynamic knowledge sharing, which may help governments and non-governmental delegations in preparing for the GGE activities in 2020, and its recommendations to the next Meeting of the High Contracting Parties, in 2020, and the Sixth Review Conference of the CCW, in 2021.

In its November 2019 meeting in Geneva, the CCW High Contracting Parties endorsed the GGE-LAWS recommendations, particularly a set of guiding principles. The recommendations affirmed that international law, in particular IHL, and relevant ethical perspectives should guide the work of the GGE. The GGE will submit two reports: one to the Meeting of the Parties in 2020 and another to the Sixth CCW Review Conference in 2021. In its discussions, the GGE will consider the legal, technological, and military aspects in an integrated manner, bearing in mind ethical considerations. According to the recommendations, the GGE will use those elements "as a basis for the clarification, consideration, and development of aspects of the normative and operational framework on emerging technologies in the area of lethal autonomous weapons systems."

The Rio Seminar was the first of a chain of similar events that took and will take place in 2020 and 2021 in both virtual and presential formats in Berlin, Geneva, Tokyo, and elsewhere, with a view to improving the situational awareness on LAWS and helping draft national, regional, and institutional contributions to the GGE, building a spirit of consensus in the preparation of its recommendations.

The Program of the Rio Seminar comprised three panels that discussed the key topics of the GGE:

Panel 1: Human-Machine Interaction and Human Control: From Engineering to IHL

Panel 2: International Law, Including IHL, on LAWS: Is There a Need for a New Protocol?

Panel 3: Strategic and Military Dimensions of Autonomous Weapons – Disruptive Technology as a Game Changer

The readers of this publication will have the pleasure of enjoying the high quality of the presentations. Furthermore, those presentations may be watched at FUNAG's portal: <http://funag.gov.br/index.php/en/news/3072-registrations-open-for-the-rio-seminar-on-autonomous-weapons-systems>. I thank each and every panelist for their engagement in the debates and the extraordinary contribution for the advancement of the complex issues at stake.

Brazil is proud to help contribute to the discussion on the governance of LAWS.

We are aware of the need not only to ensure full compliance with the IHL already established, but also to enhance international law, particularly IHL itself, with new legal and technical rules and parameters to catch up with the fast weaponization of AI technology. This is why Brazil favors a new protocol to the CCW that ensures, on the one hand, the balance between defense and security needs and technological progress and, on the other hand, the fulfillment of humanitarian purposes in the spirit of the Geneva conventions. This is one of the challenges of the present generation.

In response to the GGE Chairman's request for comments on how to operationalize the guiding principles on LAWS approved by the High Contracting Parties of the CCW, Brazil proposed two documents as national contributions to the debate. The first, ***Operationalizing the Guiding Principles: A Roadmap for the GGE on LAWS***, suggests paths of action towards a legally binding instrument (protocol on LAWS) aimed at enhancing IHL on that matter. The second, ***LAWS and Human Control: Brazilian Proposals for Working Definitions***, focuses on the concepts of LAWS itself and human-

machine interaction, particularly human control, and accountability in the employment of those systems. Both documents are presented right after this introduction.

Alessandro Candeas

Ambassador, Director of the Department of Defense
Ministry of Foreign Affairs of Brazil

WORKING PAPER 1

CCW/GGE.1/2020/WP.3

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

6 August 2020

English only

Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems
Geneva, 21-25 September and 2-6 November 2020

Operationalizing the Guiding Principles: a roadmap for the GGE on LAWS

Submitted by Brazil

Introduction

1. The Chairman of the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS), within the Convention on Certain Conventional Weapons (CCW), requested comments on the operationalization of the guiding principles[1].

2. In this context, Brazil would like to put forward four paths of action to build upon the guiding principles and fulfill the mandate of the GGE. The four paths, and its working methodology, are based on a "bottom-up" approach that benefits from domestic advancements in policies and legislations, networking of experts, multi-stakeholder approach, and international cooperation.

3. The proposal consists of four sets of initiatives that, in an integrated manner, could build synergies and confidence leading to consensual advancements in the governance of LAWS. The ultimate goal is achieving codification through specific International Humanitarian Law (IHL) rules in a legally binding instrument – a new protocol on LAWS under the CCW.

Path 1

4. Establishing links between national and international regulations, and promoting cooperation, training, and exchanges with a view to contributing to the development of domestic legislation, public policies, directives, and doctrines on LAWS, in compliance with international law, including IHL, as well as of Article 36 of the Additional Protocol I (1977) to the Geneva Conventions (1949).

5. This path would address guiding principles "c", "d", "e", "f".

6. States-parties would be encouraged to share their policies and best practices within the GGE. These domestic policies, best practices, and regulations could include national directives, normative frameworks, rules of engagement, chains of command and control, measures for accountability and transparency, requirements for designing, developing and

[1] Annex IV of the Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (CCW/GGE.1/2019/3). Available at: https://undocs.org/en/CCW/GGE.1/2019/3.

GE.20-10462(E)
2010462

acquiring AWS, security, procedures for safety and risk mitigation, including against terrorism, as well as cybersecurity against hacking and spoofing.

7. Progress and transparency on national practices and regulations will exert a positive impact on the international sphere, building confidence and a common ground for a codification endeavor.

Path 2

8. Setting up a network of legal experts, and broadening the dialogue with other UN fora. Principles addressed: “a”, “c”, “d”, “h”.

9. The proposal of an international network of legal experts on LAWS aims to enhance discussions on legal issues related to LAWS, with a view to (i) establishing the set of international law, in particular IHL, applicable to LAWS; (ii) identifying possible gaps in the normative framework in which it regards to the new challenges posed by LAWS in the following issues: accuracy in fulfilling the principles of distinction, proportionality, precaution; the prohibition of indiscriminate attacks; protection of combatants and civilians and reduction of collateral damage; accountability for rules of engagement and chain of command and control; and (iii) identifying and disseminating advancements in domestic legislation (in connection with Path 1 above).

10. Brazil suggests that the GGE/LAWS invites UNIDIR to act as a hub of the aforementioned network. The network of legal experts could submit a report to the GGE/LAWS, which could forward it for consideration by the next Review Conference of the High-Contracting parties of the CCW.

11. Brazil suggests, moreover, that the GGE/LAWS maintains a dialogue with the GGE on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, as well as the Open-Ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security.

Path 3

12. Conferences on IHL standards for the development of artificial intelligence: government, industry and other stakeholders

Principles addressed: “b”, “c”, “d”, “f”, “g”, “h”, “i”, “j”.

13. As the Rio Seminar on AWS[2] pointed out, effective regulation on LAWS may profit from other methods, besides legal texts: political declarations, corporate codes of conduct, market rules and restrictions, system architecture, programming benchmarks and shared military doctrines.

14. This path suggests the organization, with the participation of the GGE, of multi-stakeholder events and researches involving governments, the private sector, the scientific community and military experts. As with the international network of legal experts, these events could present the summary of their discussions to the GGE, addressing issues like certification requirements, the establishment of IHL benchmarks for AI engineers, machine lifecycle, market regulations, corporate codes of conduct, government acquisitions and procurements.

15. Those events and researches could dig into technical, corporate and military discussions on AWS and human-machine interaction, human control, system architecture, algorithms, syntax, the semantics of programming language, physical security, safeguards,

[2] The Rio Seminar on Autonomous Weapons Systems was held on February 20, 2020. See videos and presentations at http://www.funag.gov.br/index.php/en/news/3072-registrations-open-for-the-rio-seminar-on- autonomous-weapons-systems.

failure mode analysis, risk assessment, cybersecurity against hacking and data spoofing, mitigation measures, cyber warfare, and environments of the use of force involving AI.

Path 4

16. Promoting a strategic agenda for LAWS. Principles addressed: "c", "d", "f", "h", "j", "k".

17. This path puts forward the proposal of setting up a network of focal points from Ministries of Foreign Affairs, and Ministries and equivalent authorities of Science and Technology. The network will discuss LAWS, exchange and disseminate best practices, doctrines, and policies established by national defense strategies, white books and other documents (in connection with Path 1 above) in order to generate confidence building through convergent approaches, verification measures and to prevent unlawful proliferation, escalation, and accession by terrorist groups.

18. This path envisages the strategic discussion on LAWS within the agenda of Defense regional and multilateral mechanisms and meetings (at Summit or Ministerial levels) with a view to issuing political declarations addressing commitments to IHL compliance, improvement and accountability, together with cooperation with regard to the implementation of Article 36, above mentioned.

Towards a normative framework

19. The four paths aim to allow the GGE to profit from a multifaceted universe of perspectives from various stakeholders and at different levels and bring in their rich discussions on the challenges posed by LAWS. The guiding thread of the proposals is the operationalization of the principles "c" and "d" towards a normative framework.

20. Brazil believes that the human-machine interaction (principle "c"[3]), including human control, should be the cornerstone of the GGE debate and recommendations on LAWS governance, so as to assure compliance with international law, in particular IHL. Accountability (principle "d"[4]) is likewise a key factor to assure compliance with international law, in particular IHL, for it envisages the employment of AWS under rules of engagement and within chains of command and control.

21. Nuclear, chemical, and biological weapons were fully operational when regulatory regimes were established by legally binding instruments. In contrast, LAWS and other emerging technologies are under fast development and will keep on evolving, in parallel to the discussion on the need for specific regulations under international law. Thus, it is meaningless to wait for LAWS further development to start negotiating a legal framework.

22. The extraordinary speed of the weaponization of AI does not allow for the luxury of long years hesitating on the establishment of a normative framework.

23. The proliferation of LAWS is a risk multiplied by the very nature of self-learning machines, with relatively unpredictable behavior, in a scenario that might turn irreversibly

[3] Principle "c": Human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon, should ensure that the potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems is in compliance with applicable international law, in particular IHL. In determining the quality and extent of human-machine interaction, a range of factors should be considered including the operational context, and the characteristics and capabilities of the weapons system as a whole. See also Brazil's Working Paper on LAWS (CCW/GGE.2/2018/WP.5).

[4] Principle "d": Accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control.

out of control. Factual reality would make discussions and negotiations irrelevant after some technological thresholds are crossed.

24. Not engaging in the governance of emerging technologies in order to avoid constraints to strategic advantage capabilities is a counterproductive misperception. *Jus in bello*, in the spirit of the Geneva conventions, does not hamper strategic competition and technological development. Its purpose is to frame it in a way compatible with military necessities while protecting civilians and combatants according to humanitarian principles long approved by the international community. IHL enhancement with regard to LAWS is in the interest of collective security.

25. In view of all this, Brazil proposes initiating negotiations of a legally binding instrument on LAWS in the form of a new Protocol to the CCW, as an outcome of a collective, synergic endeavor undertake accordingly to the four paths above. Existing IHL rules are unsufficient to ensure fully responsible use of AWS, nor provide adequate means for enforcing the principles of distinction, proportionality, precaution, and protection.

26. The codification of new IHL rules could establish a balance between, on the one hand, defense and security needs and technological development without establishing asymmetries among "haves" and "have nots" and, on the other, compliance with humanitarian principles and normative.

27. A protocol could be applied to LAWS in a way compatible with evolving technology, while safeguarding the centrality of the concept of human control. It could establish a general obligation of maintaining meaningful human control over the use of force through the activation of AWS, as well as specific obligations regarding critical functions. The production and use of certain categories of AI weapons could be prohibited. Verification, compliance, transparency and enforcement mechanisms could be defined, as well as cooperation measures to help implementation on the national level. Review meetings among the contracting parties could be convened to assess the implementation of the treaty and propose, if needed, adaptations and updates.

WORKING PAPER 2

CCW/GGE.1/2020/WP.4

Convention on Prohibitions or Restrictions on theUse of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects

19 August 2020

English only

Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems

Geneva, 21-25 September and2-6 November2020

LAWS and human control: Brazilian proposals for working definitions

LAWS

1. The weaponization of Artificial Intelligence (AI) – the so-called algorithmic warfare, notably in association with robotics, cyber warfare, drone, and missile technology – has given rise to artifacts of singular nature notwithstanding century-old efforts to regulate the conduct of hostilities and the means of war. Since AI warfare has produced unique weapons, the issues raised by them must be addressed distinctively *vis à vis* conventional artifacts.

2. Autonomous Weapons Systems (AWS) are different for a set of reasons:

a) Being "intelligent," they are capable of evolving on their own, due to the functions of self-learning and self-(re)programming.This being so, they are essentially unpredictable in the long run, for some parameters imbedded in their software may be overruled and "improved" by the systems themselves. Therefore, intelligent machines may bypass human instructions and find breaches in command and control, what makes necessary theestablishment of limits at an early stage of their design and development;

b) They are more performant, and increasingly more lethal;

c) From the engineering point of view, AI inserts an upper layer of abstraction above the system's programming language; by doing so, it widens the "cognitive distance" between the decision to activate an AWS and the consequences of the attack; since AI further isolates the human operator from the "heat of the battle," the user's perception and decision-making will tend to be more abstract and detached from the intuitions and emotions that arise from close contact with enemies;

d) While the operator may receive better information on the conflict environment, certain critical functions will be outsourced to the machine during the attack procedures (tracking, targeting, locking, engaging);

e) The environment that informs the operator is mathematically modeled and may be subject to misunderstandings and malfunctions; human errors may thus be replaced by cyber misinterpretations of the environment or situational awareness, or by system biases.

3. Given the extraordinary complexity of the subject matter and the rapid pace of AI technology involved, there is still no consensus on the definition of AWS. Nevertheless, technical complexities should not hinder progress in the discussion of LAWS governance, which should be based upon the concept of human-machine interaction, particularly human control, in compliance with IHL. The "conceptual trap" may be proved counterproductive

GE.20-10872(E)

2010872

Please recycle

in the long run, for IHL enhancement with regard to LAWS is in the interest of collective security.

4. Thus, Brazil favours a workable, pragmatic definition of LAWS, that goes beyond the "technology-centric definitional approach". The concept proposed by ICRC-SIPRI[1], elegant in its simplicity, is of great usefulness in this regard, and should be adopted by the GGE:

> *"Autonomous weapon system is any weapon system that once activated can select and attack targets without human intervention."*

5. For a more comprehensive definition of AWS, Brazil proposes the following addition: "An intelligent weapon system with autonomous operation mode (i.e., without human input after activation) capable of recognizing patterns in combat environments, and of learning to operate and make decisions regarding the critical functions of target identification, tracking, locking-on and engaging based on uploaded databases, acquired experiences and its own calculations and conclusions."

Human-machine interaction and human control

6. To what extent can algorithms, syntax and semantics of the programming language of AWS comply with the principles of distinction, proportionality, precaution, prohibition of indiscriminate attacks, protection of combatants and civilians, and reduction of collateral damage in the absence of human control?

7. Who will be held accountable for the misuse or the eventual unintended result of the use of an AWS?

8. What levels of unpredictability – a key feature of IA and AWS – are acceptable to IHL?

9. The above questions put the objective notion of human control at the center of the discussion on human-machine interaction and accountability in the use of AWS. The cornerstone of the work of the GGE must be the concept of human control instead of subjective concepts like "human judgment" and "intent."

10. Human-machine interaction is the link between, on the one side, engineering, and operational system, and the other, the operator. The machine, extension of the human operator, responds to the user's consciousness, judgment, knowledge, professional training, and intent.

11. This interaction takes place in two spheres: software, including programming language and database matching; and hardware, including drones, robots, missiles, or vehicles. Both areas of interaction follow strict rules of engagement and command and control, linking the operator to his superiors in compliance with military protocols and legal rules.

12. Since AI adds an upper layer of abstraction on top of the programming language, as mentioned earlier, the programmer and the operator do not have full control over the behaviour of the machine; instead, they set goals and rules that are read by the "inference engine", allowing the machine to take its own decisions according to those parameters. Thus, autonomous systems reduce the controlling role of the programmer, and even less control is left to the operator. Human control will be increasingly challenged by the sophistication of AWS, adding higher levels of unpredictability to the behavior of intelligent warfare if limits are not put in the earlier stages of their lifecycle.

[1] SIPRI-ICRC. Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control.Available at: https://www.sipri.org/publications/2020/other-publications/limits-autonomy-weapon-systems-identifying-practical-elements-human-control-0

13. After activating the device ("fire and forget"), the operator may not be totally sure of the ultimate target, or of the time and location of the attack. Since the machine behavior may be different from the user intent, there must be some level of "human on the loop" control in order to achieve the desired result.

14. Moreover, AWS receive inputs from the environment, which also may be misinterpreted by the system or changed after the moment of activation of the system.

15. Although AWS may provide better situational awareness and tactical-operational efficiency, as well as a much more accurate and efficient response in compressed time-frames (e.g., against missiles or lasers), human control exerted by combatants is necessary to make accurate judgments in the conduct of hostilities in order to both achieve military purposes and to assure compliance with IHL. This includes the possibility of intervening to override the machine's action and terminate engagements, especially in the event of system failure.

Human control as the cornerstone

16. AWS changed the place of users from manual operators to supervisors of the machine's operations. Since intelligent machines are "logical," but not "reasonable," lacking common sense and abstract thinking, and since they reduce the controlling function of programmers and users, humans must retain the ability to supervise, intervene and deactivate attack procedures, for they possess cognitive, holistic and intuitive capabilities that AWS do not have: qualitative judgment, reasoning, andreflection about the consequences of specific attacks. Moreover, the role of human sensibility in decisionsthat cause loss of lives and the destruction of houses, buildings, and facilities, should not be overlooked. Those complex capabilities cannot be inserted into AI systems, but they are inherently present in the minds and the personal experience of commanders and combatants within the framework of war protocols, rules of engagement, chains of command and control and interpretation of IHL rules.

17. The concepts of "human judgment" and "human control" are not only compatible but necessarily interlinked. They are not mutually exclusive, for they refer to different levels of the human-machine interaction (or teaming): "human judgment" involves the doctrine of employment, while "human control" is the operation of the weapon itself. Since it is not the scope of this paper and of the GGE mandate to discuss military doctrine – the realm of "human judgment" –, the focus should be puton the operation of the AWS – thus on "human control".

18. The objective concept of "human control" refers to the human-machine interface (HMI) and the modes of operation of the weapon: Off, Stand-by, Manual, Semi-auto, and Auto.

19. On its part, the broader and subjective concept of "human judgment" refers mainly to the discernment ability of the individuals under the chain of command and control (commanders, supervisors, operators) related to the weapon deployment, taking into account the doctrine, the habilitation of the various modes of operation, rules of engagement, training, and combat contexts.

20. However, to ensure that machines execute the intent of commanders and operators in the use of force solely on the basis of human judgment is not sufficient. Accountability must be required in the case of the unintended result of the use of an AWS: for instance, a requirement for the insertion of the supervisor's password to go from Semi-auto to Full Autonomous mode of operation.

21. Given the nature of AWS, the machine behavior may cause "unintended engagements" different from the user "intent," informed by the operator's "judgment," in the absence of human control. Human control is thus the sole concept capable of assuring the responsible use of AI in weapons systems. Responsibility, accountability, and liability in the event of unlawful employment caused by intent, guilt, deceit, recklessness, negligence, or malpractice must be ensured.

22. In synthesis, lawful AWS operations must rely not on "human judgment" or "intent" – which are essentially subjective –, but on the objective concept of "human control" over the critical functions and supervision to correct autonomous decisions that produce collateral damage, override system failures or misinterpretations of the environment, target, timing and to achieve the desired outcome both in military and legal terms.

23. A responsible chain of command and control cannot outsource the compliance with IHL – distinction, proportionality, precaution – and the moral and legal implications of unlawful use of force to inanimate machines, regardless of their sophistication and intelligence. Moreover, it is essential to clarify the causal link between the agent's conduct and the violation. Good faith and adequate judgment disconnected with meaningful control may not be sufficient to assure compliance with IHL rules in the operation of intelligent machines. Deployment of AWS involves a degree of risk assessment and responsibility that cannot be free from accountability under international law and IHL.

24. The already cited ICRC-SIPRI report underlines that human control can be exercised in three ways: controls on the AWS parameters, on the environment, and on human-machine interaction. The report also examines the phases when requirements for human control may be operationalized or implemented: study, research, and development, procurement, deployment. The GGE could further elaborate on how those controls could be translated into IHL parameters.

25. The discussion on human control should take into account defensive and offensive actions.

26. In a defensive scenario, given the lack of time to respond to missile attacks, for example, and in the interest of protecting combatants and especially civilians, some of the critical functions must be done autonomously. In these situations, greater flexibility may be granted to AWS.

27. On the other hand, at offensive scenarios, greater levels of human control and limited autonomy on critical functions should be mandatory in combat situations with deployment of AWS, especially in populated areas.

Working definitions for a legal framework

28. This paper is linked to the other Brazilian contribution to the GGE presented under the title *Operationalizing the Guiding Principles: a roadmap for the GGE on LAWS* [2].

29. In that document, Brazil proposes paths of action leading to advancements in the governance of LAWS, ultimately arriving at the codification of specific International Humanitarian Law rules and a new protocol under the CCW.

30. Such a protocol could establish the general obligation of maintaining meaningful human control over the use of force through the activation of AWS, as well as specific obligations regarding control over critical functions of selecting and engaging targets. Furthermore, specific categories of AI weapons should be prohibited on the basis of ethical and moral considerations.

31. The working definitions presented in this paper are designed to contribute to the drafting of that legal framework.

[2] Available at: https://documents.unoda.org/wp-content/uploads/2020/08/CCW-GGE.1-2020-WP.3-.pdf

Rio Seminar on Autonomous Weapons Systems – Concept Note

Rio Seminar on Autonomous Weapons Systems
Rio de Janeiro, Naval War College
February 20, 2020

Objectives of the event:

Contributing to the debate on the governance of the emerging technologies in the field of LAWS in the scope of International Law, including Humanitarian International Law, particularly in the framework of the GGE/LAWS of the CCW, considering its diplomatic, legal, technological, business, strategic, and military dimensions.

The Seminar aims to promote discussions in an informal environment, with the purpose of enabling an improved exchange of information, which can assist governments and non-governmental delegations in preparing for the GGE/LAWS in 2020 and in drafting their recommendations for the next CCW Meeting of the High Contracting Parties, in 2020, and for the 6th CCW Review Conference, in 2021.

Participants:

Representatives of governments, international organizations, the International Committee of the Red Cross, NGOs, private enterprises, and academia.

Background:

The High Contracting Parties to the CCW endorsed the recommendations of the GGE/LAWS, particularly a set of guiding principles, in their meeting in November 2019, in Geneva.

Among other aspects, the recommendations confirmed that the work of the GGE must be guided by International Law, in particular Humanitarian International Law, and by ethical aspects.

The GGE will submit two reports: one for the meeting of the Parties in 2020 and another for the 6th CCW Review Conference in 2021.

In its discussions, the GGE will consider the legal, technological, and military aspects in an integrated manner, taking into account ethical considerations.

According to the recommendations, the GGE will utilize the aforementioned aspects "as a basis for the enlightenment, consideration, and development of aspects of the normative and operational structure on emerging technologies in the field of lethal autonomous weapons systems."

It is in this context that the Seminar is inserted.

The event will allow for informal and high-level interaction between governmental and non-governmental participants, with the objective of sharing knowledge and improving situational awareness on LAWS. The discussion can aid in the drafting of national and institutional contributions for the GGE/LAWS and in building a spirit of consensus in the preparation of recommendations.

Program

Thursday, February 20

09:00 – Opening session

Roberto Goidanich, Minister, President of FUNAG

Alvaro Monteiro, Squadron Admiral, President of the Center for Political-Strategic Studies of the Brazilian Navy

Merel Ekelhof, UNIDIR

Janis Karklins, Ambassador, Latvian Mission to the United Nations in Geneva, Chairman of the 2020 GGE/LAWS

Alessandro Candeas, Ambassador, Director of the Department of Defense of the Brazilian Ministry of Foreign Affairs

09:40 – Coffee break

10:00 – Panel 1: Human-Machine Interaction and Human Control: From Engineering to IHL

Amanda Wall, Office of the Legal Adviser – Department of State (United States)

Elizabeth Minor, Article 36 / Campaign to Stop Killer Robots

Yokoyama Daiki, Conventional Weapons Division – Ministry of Foreign Affairs (Japan)

Anja Dahlmann, German Institute for International and Security Affairs (Germany)

Geber Ramalho, Cesar Institute (Brazil)

Edson Prestes, Federal University of Rio Grande do Sul (Brazil) – Moderator

12:30 – Lunch break

14:00 – Panel 2: International Law, Including IHL, on LAWS: Is There a Need for a New Protocol?

Konstantin Vorontsov, Ministry of Foreign Affairs (Russia)

Kathleen Lawand, Legal Division – Arms Unit (Red Cross – ICRC)

Michael Biontino, Ambassador, Special Adviser – Foreign Office (Germany)

Bonnie Docherty, Harvard Law School

Thomas Hajnoczi, Ambassador, Director of the Department of Disarmament, Arms Control and Non-Proliferation – Ministry for Europe, Integration and Foreign Affairs (Austria)

Pamela Moranga Quezada, Delegation to the United Nations in Geneva (Chile) – Moderator

16:30 – Coffee break

16:45 – Panel 3: Military Dimensions of Autonomous Weapons – Disruptive Technology as a Game Changer

Karl Chang, Associate General Counsel, Department of Defense (United States)

Roberto Gallo, ABIMDE (Brazilian Association of Security and Defense Industries)

Chen Yongcan, Deputy Consul General (China)

Moa Peldán Carlsson, Stockholm International Peace Research Institute (SIPRI)

Alfredo Muradas, Vice-Admiral, Director of Weapons Systems, Brazilian Navy (Brazil)

Antonio Jorge Ramalho, University of Brasilia (Brazil) – Moderator

18:45 – Closing session

Opening Session

It is with great satisfaction that we note the presence of Squadron Admiral (Marine) Alvaro Augusto Dias Monteiro, President of the Center for Political and Strategic Studies of the Navy (CEPE-MB);

Of Squadron Admiral José Antônio de Castro Leal, Counsellor at the CEPE-MB;

Of Ambassador Janis Karklins, Chairman of the 2020 GGE on LAWS;

Of Ambassador Alessandro Warley Candeas, Director of the Department of Defense of the Ministry of Foreign Affairs;

Of Minister Roberto Goidanich, President of FUNAG, and other officials present here.

The Ministry of Foreign Affairs and the Alexandre de Gusmão Foundation, jointly with the Center for Political and Strategic Studies

of the Navy, the Brazilian Naval War College, has the honor to welcome the panelists and members of delegations present here for the international Rio Seminar on Autonomous Weapons Systems.

We invite Minister Roberto Goidanich, Ambassador Alessandro Candeas, Squadron Admiral (Marine) Monteiro, Ambassador Janis Karklins, and UN researcher Merel Ekelhof to make up the opening session panel.

Squadron Admiral Alvaro Monteiro

Counsellor at the CEPE-MB

Ladies and gentlemen, good morning to everyone!

I hope you have been, in tandem with your professional activities, able to enjoy the beauties of the festive environment of this beautiful city of Rio de Janeiro.

We cannot begin this seminar without first registering our great satisfaction in participating in such a relevant, such an important event for all those interested in the theme at hand, the Lethal Autonomous Weapons Systems.

As humanity breaks its boundaries in knowledge, it glimpses new prospects for human life, which it tries to reach in a ceaseless technological rush. Among these prospects, it is important to highlight artificial intelligence. Artificial intelligence has been developing with

uncommon intensity, unfolding in diverse segments, such as systems capable of learning through the automation of the construction of analytical models, the well-known machine learning, or machines capable of learning through neural networks, deepening their initial knowledge, gradually adjusting their parameters until obtaining the desired results, a process better known as deep learning.

The Armed Forces are no exception in the absorption of these technological innovations; they seek to incorporate them with maximum efficiency into their weapons systems. It is clear that artificial intelligence can increase the accuracy of the deployment of weapons systems and reduce expenditure in military resources, especially human ones.

However, the relative autonomy these systems may come to incorporate, if they have not already, making decisions with minimal or no human intervention, no longer necessarily following the decision-making processes preprogrammed by the humans who built them, brings concerns and restlessness to humanity.

The deployment of lethal autonomous weapons, the LAWS, is already a reality in some specific cases. However, the technical uncertainties as to the behavior of such systems and the legal and moral dilemmas stemming from their development and occasional deployment have raised new questions, including about the ethical dictates that modulate the preprogrammed algorithms in this system. How can one be sure that human ethics was correctly incorporated into them? And what would effectively be human ethics? Does it have a universally consecrated value? Or would it vary in accordance with human ambiguities? This is because, metaphorically speaking, human beings are also guided by algorithms. As well put by Émile Durkheim, father of sociology as a science, we are the product of our own circumstances, of the social facts that exert their coercive

power over us. Social circumstances that are introjected into us throughout our existence, some even before we arise into life.

Consequently, even if we deem ourselves capable of thinking and acting critically based on universally consecrated moral values, depending on our sensitivity, the initial conditions of a given process, we are liable to make choices that we had previously never thought possible. Then there are the questions that cause humanity disquiet: does it make sense to develop lethal weapons systems that do not submit to human control? Does it make sense to develop lethal weapons systems whose action will depend exclusively on the imponderability of the ethical, political, and military biases they have inferred from their respective algorithms?

It is in the attempt to answer such questions that humanity realizes that, parallel to the quick development of disruptive technologies, it must also seek to previously develop, with equal intensity and effort, legal rules and principles that regulate their use, or perfect the existing ones so that they conform to new technological developments.

A task, however, that is not easy. Quite the opposite, it may be even more complex than the very development of the systems it intends to regulate. This happens because, apart from the difficulty in normatizing, regulating situations that are still not perfectly defined and universally accepted, there are the great discrepancies between the technological levels of states, which cause them to resist rules they think may come to characterize technological curtailment. That is why there is a difficulty in reaching a common denominator capable of amalgamating the several conflicting interests—a circumstance that, by delaying the establishment of such regulations, only highly increases the need for their existence.

From there comes the greatest reason for this seminar: to attempt, through the expositions, the debates, and the talks among

all the participants—in an informal and free environment, one of the dictates assumed by the states involved—to try to reach a common basis, a consensual view that can facilitate understanding in future forums and conventions on the theme. In opening the activities of this seminar, I can only wish that we all have a productive and beneficial day, and that we may succeed. Thank you very much.

Minister Roberto Goidanich

President of FUNAG

Thank you very much. Good morning, ladies and gentlemen.

I would just like to make a few brief acknowledgments.

Firstly, Squadron Admiral Alvaro Monteiro, President of the Center for Political and Strategic Studies of the Navy, and the other admirals present here. The Naval War College has been an excellent partner to the Alexandre de Gusmão Foundation. This is already the second event in little over half a year. The last event was conducted here in August last year, precisely on the same theme of LAWS. It was a business and academic symposium on autonomous weapons. We even have a few videos we made about that event, they are all available on FUNAG's YouTube channel.

So we profusely thank the Naval War College for this reception, for welcoming us into this wonderful auditorium. My deepest thanks.

Obviously I also thank the Brazilian Ministry of Foreign Affairs, in the person of Ambassador Alessandro Candeas, Director of the Defense Department in the Ministry of Foreign Affairs, who is also a traditional partner to the Alexandre de Gusmão Foundation. We conducted several other events with him last year.

I also thank the Department of the United Nations of the Brazilian Ministry of Foreign Affairs, in the person of First Secretary Daniele Farias Luz. We recently held an interesting event, in late October 2019, which also addressed this theme of LAWS. It was a commemorative event of the 70 years of the Geneva Conventions, about current challenges to the application of International Humanitarian Law, and the last panel, specifically, dealt with these new technologies used in war, and with how International Humanitarian Law could be applicable to all these new technologies. I would even invite anyone interested in these presentations to watch the videos on FUNAG's YouTube channel, or listen to the podcasts (FUNAG's podcasts are available in ten different platforms, such as Apple Podcasts, Spotify, Google Podcast, Anchor). So, for those who are interested, I highly recommend these presentations, because it was also a very interesting discussion, the one we had in late October.

I also thank Ambassador Gonçalo Mourão for his presence. He is Brazil's Permanent Representative for Disarmament in Geneva.

And I obviously thank very much our foreign participants, in the person of Ambassador Janis Karklins, Ambassador of the Latvian Mission to the United Nations, and Chairman of the Group of Governmental Experts on LAWS, who dared to walk here from the hotel in the Rio de Janeiro heat. It is really impressive!

I also thank Merel Ekelhof for her presence. She is from the United Nations Institute for Disarmament Research (UNIDIR).

And to everyone else who came from abroad, I would like to extend our warmest welcome to Brazil, and very special thanks to our foreign participants from Germany, from the United States, from Russia, from Austria, from Japan... Yesterday I talked to our colleagues from Japan, who made a very extended trip, a 24-hour trip from Tokyo to Rio just to participate here in this seminar. Thank you very much! I really appreciate it.

I also thank our Brazilian researcher colleagues, who also honor us with their presence: Antônio Jorge Ramalho, from the University of Brasília; Edson Prestes, from the Federal University of Rio Grande do Sul; and Geber Ramalho, from the Federal University of Pernambuco. Thank you very much for honoring this event.

And also, obviously, all of the ladies and gentlemen present here in this auditorium, who will enjoy this day of very interesting debates on a new topic, which is LAWS.

I also thank those who are watching us through the Internet. We are broadcasting this event live.

Now, we must explain this to the Brazilians watching us: we will broadcast with simultaneous translation into English, and the English presentations will not be translated into Portuguese during the live broadcast. But, later on, in the next few days, we will produce videos with interpretation into Portuguese, so that, later, everyone can have access, on FUNAG's YouTube channel, to all of the events with simultaneous translation into Portuguese.

And, obviously, I thank, as usual, our staff at FUNAG, without whom this would not have been possible. They have done extraordinary work here in the organization of the event, as well as our outsourced help, the Brazilian Sign Language interpreters, the simultaneous translators. Thank you very much!

We also intend to publish a book on the results of this seminar, on the presentations that will be made throughout the day. Ambassador

Alessandro Candeas already told the participants about this idea yesterday. So we would like, insofar as possible, that those of you who wish to do so later please send your texts, your contributions, to Ambassador Candeas, so that we can later compile a publication by the Foundation, which will also be available for free download from our digital library.

While on the subject, I can never neglect mentioning, our digital library has close to 800 volumes for free download. So I recommend that all who are interested in themes of foreign policy and international relations visit our digital library. It has actually had millions of downloads, and I think it is a very rich material on those themes. I hope this event also contributes to further enrich our library.

So these were my brief thanks. I will not take up any more time from the event, but thank you very much, especially to the Naval War College for having us.

Merel Ekelhof

UNIDIR

Excellencies, ladies and gentlemen,

Let me begin by expressing my thanks for the opportunity to address you today. By way of a brief self-introduction, my name is Merel Ekelhof, I am a researcher at UNIDIR, leading the AI and Autonomy Portfolio of the Security and Technology Program. A central aim of UNIDIR's work on Lethal Autonomous Weapons Systems in 2020 is increasing the understanding and operationalization of concepts like "human control" and "human-machine interaction" (HMI). We have selected this topic given the importance that this subject will have in the context of the development of the normative and operational framework that the Group of Governmental Experts is mandated to consider in 2020 and 2021. Now that states have

agreed on the 11 Guiding Principles, including one specifically on human-machine interaction, it is time to examine how these principles are operationalized in military practice. This year, UNIDIR intends to organize a number of table-top exercises—including one in this region—to guide deeper discussion on the military and legal aspects of human control.

Having said that, I would like to take this opportunity today to discuss **three fundamental questions** that we think are of critical importance when we discuss concepts like human control or HMI.

1: What do we ultimately want to achieve? HMI or human control ***to what end***?

2: HMI or human control ***over what***?

3: HMI or human control exercised ***by whom***?

1: What do we ultimately want to achieve? HMI or human control to ***what end***?

It seems that discussions on HMI are complicated by diverging motivations. What is the ultimate aim of having a certain standard of human control or HMI? While Lethal Autonomous Weapons Systems are not regulated expressly by any treaties, it is undisputed that international law applies to both the development and use of LAWS.

The prevailing argument in CCW discussions on LAWS is therefore that HMI should be aimed at ensuring compliance with existing applicable law. It is most regularly argued that the need for human involvement flows from the desire to ensure legal compliance. Nonetheless, according to some, HMI would preferably satisfy more than compliance with legal obligations, and should also include moral and ethical standards.

It remains controversial whether HMI should be aimed at complying with existing law or prescribing a higher standard. Before we can move into more detail of how human control or HMI should

be operationalized and whether new regulation is needed, it is important to have a clear (and preferably shared) understanding of what states want the principle to satisfy.

2: HMI or human control ***over what***?

Considering that HMI is a principle introduced and discussed in the context of LAWS debates, it is often assumed that human control should be exercised over LAWS. This control can thus be exercised over LAWS as a specific category of weapons, but it may also apply to weapons in general, parts of the weapon (such as critical functions) or, more broadly, over the life cycle of the weapon.

However, human control has also been framed in relation to a range of decisions, actions, or a process. Some examples are: human control over every individual attack, the targeting process, the use of force or, more narrowly, the final decision to use force.

To gain a better understanding of how HMI and human control should be operationalized, it is, thus, important to understand: 1) ***what*** the principle is supposed to achieve; 2) ***over what*** part of which weapons, actions, decisions or processes the principle should be applied; and, lastly 3) who should be involved in implementing the principle. This brings us to the third question.

3: HMI or human control exercised ***by whom***?

It is not uncommon for discussions about HMI to focus on the relationship between an operator and the weapon during weapon deployment. This relatively narrow focus can be potentially problematic, because, in practice, humans may exercise different forms of control at various junctures in the decision-making cycle and at various command levels in the organizational structure. As such, the control that may be exercised during weapon deployment is only part of the picture. Without due consideration of this practical,

organizational context within which human control is typically distributed, it seems impossible to create an accurate picture of how human control or HMI can (or cannot) facilitate and ensure compliance with the applicable law.

The question here is not simply one of legal responsibility: a ***post-deployment assessment*** to assign responsibility to an individual for violating legal obligations. But includes a ***pre-deployment, preventive approach*** focusing on how one can organize military decision-making in a way that allows for appropriate human involvement to improve compliance with legal obligations throughout the process and at all levels of decision-making.

To conclude, the guiding principle of HMI and conceptualizations of human control are certainly appealing standards that are worth pursuing—as an informal guideline, as part of a political declaration or a legally binding instrument. However, rather than prejudging discussions by focusing on one of these outcomes, thereby ignoring potential others, it is important to further clarify the scope and practical application of the concept itself. The three questions provided—HC to what end? HC exercised over what? And by whom?—may be useful tools to guide these deliberations.

Ambassador Janis Karklins

Chairman of the 2020 GGE on LAWS

Thank you very much. Let me start by—on behalf of all participants of the seminar—thanking the Brazilian host, the Ministry of Foreign Affairs, as well as the Brazilian Naval War College, for hosting us and making possible this conversation throughout the day on the topic that is very close to our heart. So, thank you for that.

Secondly, let me apologize for my appearance; in the morning I was hoping to challenge the nature of Brazil, and I happened to be on the losing side. There is always a silver lining on every cloud. I will be cooled throughout the day, because my shirt is really wet.

On a more serious note, let me say that I am not a magician, and I cannot make countries change their positions if they do not

want that. So, what I would like to hope is that I can try to facilitate the process where countries would decide to change their positions by listening to others and going through the conversation, which is inclusive, even if very complex.

As a facilitator, I have given it some thought, and I would like to share with you maybe three elements of my thinking that I am now consulting with the High Contracting Parties of the CCW, to see whether that approach would be acceptable to all and if we could proceed in preparations of the GGE-LAWS Meetings.

The first element is to be rational, or as rational as we can, and try to waste neither time nor effort in doing things which would not contribute to the final result of our exercise. With that I mean that we have a two-year mandate, and I am trying to work out a process which would be stretched over a period of two years, and whatever we do in 2020 would be continued in 2021, that no effort of 2020 would remain on the shelf, but rather it would be continued. In that respect, I think that the outcome of the 2020 GGE-LAWS exercise would be in the form of a simple procedural report, which would have an annex containing the Draft Final Report of the GGE of 2021. And in that Draft Final Report we would have a structure, we would have agreed elements from the previous years and every element that we will be able to agree during our conversation in 2020, and we would put some placeholders on the topics or issues that need to be further addressed in 2021. By approving the procedural report, we would also approve a notion that the Draft Final Report would be used as a basis for further work in 2021. So, that is the first element.

The second element is to be imaginative and find a way to make progress in a very complex environment. In our work, words matter. We need to find the words that would allow us to progress without making any anticipated decisions that need to be made at the end of the process. I am covering, also, in Geneva, not only disarmament

issues, but also human rights. When I was thinking about whether there is any way or example that we could think of, I came to the human rights treaty bodies as an example. These are bodies that have been elected by member states, and they are considered the guardians of human rights treaties; they engage with the member states in examining how member states implement their specific human rights treaties, but sometimes they write commentaries. Commentaries about the implementation of certain provisions of treaties. And these commentaries are then distributed for the benefit of member states. And I was thinking whether, in our case, when we have agreed on eleven principles, or guiding principles, wouldn't it be worth inviting member states to write the commentaries on the operationalization of those eleven principles at the national level? By having this reflection on the operationalization of guiding principles, we would potentially see the emergence of commonalities of approach in operationalization, and that would probably give us some guidance for the possible framework going forward. And so far my consultations, bilateral consultations, have been reasonably productive in that respect, and I am hopeful that approach of commentaries on the operationalization of guiding principles would be one of the ways forward in addressing substantive topics of the LAWS.

Thirdly, I think we should learn from mistakes of the past and try to avoid repetition of those mistakes. In 2008, if I am not mistaken, the GGE on cluster munitions concluded its activities without result. But regrettably, we do not see any trace, in the CCW, of a discussion on cluster munitions after the failure of the GGE process. I am not suggesting, or not setting us up for failure, but I think that the topic of LAWS is so important and will be present, whether we like it or not, in [the] near and also distant future, that we cannot afford the luxury of putting LAWS aside from the agenda

of the GGE, no matter what the outcome of our conversation will be in 2021.

So these are three elements of my thinking, and I would be very happy to engage bilaterally with the participants of the seminar and see whether any fine tuning of this approach would be needed. So, with this, I would like to thank you once again for this opportunity to be in Rio, and I am looking forward to our conversation. Thank you.

Ambassador Alessandro Candeas

Director of the Department of Defense of the Ministry of Foreign Affairs

Thank you very much; I would like to begin by thanking Squadron Admiral Alvaro Monteiro, President of the Center for Political and Strategic Studies of the Naval War College, Minister Roberto Goidanich, President of the Alexandre de Gusmão Foundation, for the extraordinary support in the Rio Seminar on Autonomous Weapons Systems.

I also salute Ambassador Karklins, Chairman of the GGE, our colleague Merel from UNIDIR, Admirals, Ambassador Gonçalo Mourão, the members of the GGE present here, and our panelists, ladies and gentlemen.

If you allow me, I would like to switch now to English, saying that the main purpose of our seminar is to contribute to the consensus-building, to a better understanding of each other's positions. I would call it cognitive consonance, the search for a common ground, for a common understanding with a view to strengthening the work of the GGE and that of the High Contracting Parties of the CCW towards the international governance of LAWS by the improvement of international law, especially Humanitarian Law. What are the underlying assumptions of our seminar?

First of all, we are dealing with weapons of a different nature. Those are intelligent machines, intelligent systems capable of self-learning, self-programming, deep learning with neural networks. So the logic, the timing, and the format of the debate and regulation must also be different from that of other non-proliferation and regulation exercises in other areas of arms control and disarmament, such as nuclear, missiles, chemical, biological weapons, land mines. Those weapons, conventional weapons, already existed and were deployed in combat fields before being controlled or regulated. In contrast, LAWS will be regulated parallel to its own creation and development, trying to catch up with the fast pace of technology, and they have not yet been deployed in armed conflicts, luckily.

So traditional concepts like deterrence, attribution, verification, non-proliferation, and banning cannot be automatically employed, or may simply be useless. In particular, the logics of deterrence and wait-and-see do not apply as they do to other conventional weapons. LAWS evolve rapidly and may take over and become out of control in real combat situations. So if we adopt the wait-and-see behavior, when we finally see it, it will be too late.

There is a close connection between LAWS and cyberwarfare. Systems can be hacked, firewalls can be broken, and friends or civilians can digitally become foes in malfunctions or in changes of lines in the algorithms.

This seminar takes into account several dimensions, strategic, military, operational, ethical, legal, humanitarian, technological, entrepreneurial. We adopt here a multiple stakeholder approach involving representatives from governments, from the military, international organizations, NGOs, academia, think tanks, and also the private sector.

Our purpose here is not to rule out autonomous weapons, but to focus on using jus in bello, our focus is on jus in bello.

In the spirit of the Geneva conventions and protocols, the LAWS debate does not interfere in the warfare strategy. The LAWS debate does not impede the use of force, nor the search for strategic advantages. The LAWS debate is aimed at framing the use of force in compliance with IHL. IHL rules are in their interest not only for humanitarian purposes; IHL also meets both military and corporate concerns. We are aware that surprise has a role in military strategy itself, but also, in the same way, predictability of the rules of the game is also fundamental in strategy. Law is fundamental in military strategy, too.

LAWS must be inserted into doctrines, rules of engagement, command and control in a way that assures compliance with IHL. This brings me to a fundamental issue. The decision to kill must involve human judgment and control over the critical functions of targeting and engaging. This raises the paramount question of conscience. Human conscience and judgment cannot be simulated by algorithms, nor can the complex implementation of humanitarian principles in the battlefield be interpreted or replaced by the "if-then" strict logic embedded in the algorithms. IHL cannot be part of the semantics of the programming language. The principles of distinction, precaution, proportionality, cannot be part of the cyber syntax and semantics followed by machines. They call for meaningful

human control and judgment. International law, particularly IHL, needs enhancement, needs higher standards, as Merel just said.

Existing rules are not enough. They are somewhat vague in the face of emerging technologies; there are gaps in IHL in the governance of LAWS. But this is not only a LAWS problem; this is part of a bigger picture, a broader phenomenon. The means of war of the 21st century, the new weapons, the weapons of the war of the future, that already exist, they all lack regulation. I am talking here not only of LAWS, but also of cyberwarfare, hypersonic missiles, weapons in outer space; those weapons are not yet governed by international law.

For all these reasons, Brazil supports the beginning of negotiations to establish a positive obligation of meaningful human control in critical functions of autonomous weapons systems. Brazil favors a new protocol, a legally binding protocol, as well as other means like political declaration, corporate codes of conduct that ensure, on the one hand, the balance between the need for defense and security, strategic advantages, and technological progress, and on the other hand, the fulfillment of humanitarian purposes. We cannot afford the luxury of taking long years to establish a normative and operational framework for LAWS. At the end of this present decade, if we do not reach a consensual outcome on this, there will be no longer need for any debate. It will have become irrelevant, overwhelmed by the pace of technology itself.

Those are the points that will be raised in our discussions today; as you know, we are going to organize three panels: Panel 1: Human-Machine Interaction and Human Control: From Engineering to IHL; Panel 2: International Law, Including IHL, on LAWS: Is there a Need for a New Protocol?; and Panel 3: Strategic and Military Dimensions of Autonomous Weapons – Disruptive Technology as a Game Changer.

I wish us all a very productive Seminar. Thank you.

Panel 1:
Human-Machine Interaction and Human Control: From Engineering to IHL

Moderator: Edson Prestes

Federal University of Rio Grande do Sul (Brazil)

Excellencies, ladies and gentlemen,

Good morning.

It is a great pleasure to be here.

I am Edson Prestes, I am a professor at the Federal University of Rio Grande do Sul, and I will be the moderator of the Panel "Human-machine interaction and human-control: from engineering to international humanitarian law."

We have here, together with me, panelists Amanda Wall, Attorney-Adviser for Political and Military Affairs at the Office of the Legal Adviser of the U.S. Department of State.

We have Elizabeth Minor, Adviser at Article 36 [a non-governmental organization from the United Kingdom].

We have Mr. Yokoyama Daiki, Assistant Director at the Conventional Arms Division, Ministry of Foreign Affairs, Japan.

Anja Dahlmann, expert scientist at the German Institute for International and Security Affairs, and in charge of the International Panel on the Regulation of Autonomous Weapons, from Germany.

We have also Geber Ramalho, professor at the Federal University of Rio Grande do Sul, and President of the council at CESAR, a Brazilian innovation institute.

In this panel, we will discuss a lot of topics that include human control, human-machine interaction, robotics, and artificial intelligence.

This panel will have two parts. The first part will be the presentation by each speaker, and, at the end of the panel, we will open for questions, and of course, answers.

I would like to invite, I will give the floor to Ms. Amanda Wall for her presentation. Thank you.

LAWS and Human-Machine Interaction

Amanda Wall
U.S. Department of State

Thank you very much. Let me start by thanking our hosts, the Government of Brazil, for organizing this very important conference, the Brazilian Naval War College, for this beautiful venue and for inviting me to speak today, as well as for their leadership on this very important issue. My name is Amanda Wall, I am an Attorney-Adviser in the Office of the Legal Adviser at the U.S. Department of State, where my practice focuses on International Humanitarian Law. I am a member of the U.S. delegation to meetings of the High Contracting Parties to the CCW and the GGE on emerging technologies in the area of Lethal Autonomous Weapons Systems, and have participated in the last several meetings in that capacity.

I am going to focus my presentation today on the views of the United States on human-machine interaction in the area of LAWS.

This has been a topic that has been discussed at some length at the GGE, with a wide diversity of views presented, but having observed those debates personally, my own view is that there is actually a lot more common ground between the positions being expressed at the GGE, and the differences are not so vast as they might seem. In fact, I would posit that the states participating at the GGE actually have a lot more in common than they realize.

Let me also say, before I begin, that the United States remains fully supportive of the work of the GGE and hopes to help to accomplish a strong, substantive outcome by the end of its current two-year mandate in 2021.

In my time today, first, I am going to talk about how the United States understands human-machine interaction in the area of emerging technologies in LAWS. Second, I am going to spend a few minutes talking about why we think that the concept of "human control," as some have characterized it, does not fully capture the range of considerations that need to be undertaken when developing policies and programs for responsible development and use of LAWS, and why we think it actually may be an oversimplification of some of the very complicated issues at stake. And, finally, I am going to say a few words about what, we would argue, should be the focus of the GGE in this regard over the next two years. My remarks today draw heavily from working papers and views that the U.S. delegation has presented to the GGE, and, in particular, a U.S. working paper on human-machine interaction.[1]

But to begin, I want to make one point as a legal matter.

1 United States, Working Paper, Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Aug. 28, 2018, CCW/GGE.2/2018/WP.4, available at: <https://undocs.org/en/CCW/GGE.2/2018/WP.4f>.

Guidance from the fundamental principles of the law of war

As Ambassador Candeas noted in his introduction this morning, it is important to acknowledge first and foremost that IHL, including the fundamental principles and rules of distinction, proportionality, military necessity, and precautions in attack, continue to apply regardless of what type of weapon is used. This is reflected in the GGE's Guiding Principle (a), as well as in the GGE's conclusion in paragraph 17(a) of the GGE's 2019 report.[2]

In addition to helping assess whether a new weapon falls under a legal prohibition or how IHL requirements apply, the fundamental principles of international humanitarian law may also serve as a guide in answering novel ethical or policy questions in human-machine interaction that are presented by these emerging technologies. For example, it may be appropriate to consider the following:

- Does military necessity justify developing or using this new technology?
- Under the principle of humanity, does the use of this new technology reduce unnecessary suffering?
- Are there ways that this new technology can enhance the ability to distinguish between civilians and combatants?
- And, under the principle of proportionality, has sufficient care been taken to avoid creating unreasonable or excessive incidental effects?

2 Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Sept. 25, 2019, U.N. Doc. CCW/GGE.1/2019/3, Annex IV, p. 13, Guiding principle (a) ("International humanitarian law continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems"); Id. at par. 17(a) ("The potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems must be conducted in accordance with applicable international law, in particular IHL and its requirements and principles, including inter alia distinction, proportionality and precautions in attack").

There has been broad consensus at the GGE that IHL is applicable to the use of force, including the use of force that is reliant on autonomy or autonomous features and functions. But how do we go about ensuring that the law, in particular IHL, is complied with in the use of these weapons, and how do we develop good practices for responsible development and use of these weapons?

Effectuating the intent of commanders and operators

A big part of the answer to this question rests in human-machine interaction, which is what I have been asked to discuss today. From the U.S. perspective, the key issue for human-machine interaction in emerging technologies in the area of LAWS is ensuring that machines effectuate the intent of commanders and operators of the weapons systems. Weapons that do what commanders and operators intend them to do can give effect to their specific intentions to conduct operations in compliance with IHL and to minimize harm to civilians and civilian objects.

Much of the U.S. policy and practice in this area is laid out in a 2012 Policy Directive issued by our Department of Defense. It is DoD Directive 3000.09, titled "Autonomy in Weapon Systems." I have brought some copies of that with me, for those of you who are not familiar with it, and I will refer to it a couple of times throughout the rest of this presentation.

Let me elaborate upon the concept of good practices for effectuating the intent of commanders and operators. It is not necessarily about having a human executing or controlling every step of a weapon's operation manually: this does not happen with many weapons systems that have been in operation for decades. Instead, it is about taking practical steps that, among other things, enable personnel to exercise judgment over the use of force in an armed conflict, and reduce the risk of unintended combat engagements.

So, how do we develop and use weapons in a way that ensures that weapons give effect to human intent? What measures should we take?

Measures to ensure the use of autonomy in weapon systems effectuates human intentions

First, we need to think about how to minimize the probability and consequences of failures in weapon systems that could lead to engagements that the commander and operators did not intend. This could happen either because a weapon engaged a target that was not the intended target, or because it created unacceptable levels of collateral damage. I would note that even an attack against previously authorized targets could ultimately be "unintended" if there are significant changes in circumstances between the time of authorization and when the weapon engaged targets, such that the authorizing official no longer intended the targets to be engaged. So there is a temporal aspect to it, as well.

There are a number of ways that guidelines in the development and use of weapon systems can help minimize the probability and consequences of failures in weapon systems—failures that could lead to unintended engagements. One example is that an autonomous system might be programmed to operate only within specific geographic boundaries. If deployed and limited to an area that was a military objective, like an enemy military headquarters complex, then its use would be analogous to the use of other weapons, like artillery, that are often used to target areas of land that qualify as military objectives. Another example might be an autonomous weapon that is equipped with sensors that are designed to detect specific "signatures"—or unique, identifying characteristics that would be specific to a military objective, like frequencies of electromagnetic radiation that are generally not found in nature or among civilian objects. Many states have already used weapons that detect the specific electromagnetic signals emitted by enemy radar to help

ensure that a target is a military objective. These are all examples of ways that weapons can be developed and used to help effectuate the intent of a commander or operator by minimizing the probability of unintended engagements and minimizing the consequences of such engagements if they occur.

Second, we need to think about how to help ensure that weapon systems function as anticipated. This includes engineering weapon systems to perform reliably. The DoD Directive that I mentioned before puts in place requirements for verification and validation, and for testing and evaluation of hardware and software to make sure that they function as anticipated. For example, before fielding weapon systems that would use autonomy in novel ways, those reviews must "assess system performance, capability, reliability, effectiveness, and suitability under realistic conditions." The Directive also requires "safeties, anti-tamper mechanisms, and information assurance" to ensure that the weapon functions as it was anticipated to function, namely by helping address and minimize the probability or consequence of failures that could lead to unintended engagements or to loss of control of the system, by adversaries or others.

Third, we need to think about how to help ensure that personnel properly understand the weapon systems. This includes training personnel, establishing clear human-machine interfaces, and developing clear doctrine and procedures for use. Studies of accidents involving human use of automation have shown that failures can often result from operator error, and that better training and adherence to established tactics, techniques, and procedures and doctrine could have prevented those mistakes. That is why the DoD Directive generally requires the establishment of such "[t]raining, doctrine, and tactics, techniques, and procedures"—what we call TTPs. And, before systems that employ autonomy in new ways are fielded, senior officials must determine that "[a]dequate training, TTPs, and doctrine are available, periodically reviewed, and used by

system operators and commanders to understand the functioning, the capabilities, and the limitations of the system's autonomy in realistic operational conditions."

Further to this end, the interface between humans and machines should be clear "[i]n order for operators to make informed and appropriate decisions in engaging targets." This is why the DoD Directive requires the interface between people and machines for autonomous and semi-autonomous weapon systems to:

(a) Be readily understandable to trained operators;

(b) Provide traceable feedback on system status; and

(c) Provide clear procedures for trained operators to activate and deactivate system functions.

These are just some examples; there are a number of measures that can be taken to help ensure that weapon systems that use autonomy are developed and used to effectuate human intention in the use of force. These measures are outlined in greater detail in the working paper referenced above, which I have provided to this conference and is also available on the CCW's website.[3]

Why not "human control"?

One question you may be asking is, why not call this "human control"? After outlining the policies that our own Department of Defense has in place with regard to the use of autonomy in weapon systems, my hope is that people in the audience are thinking, well, that sounds a lot like the measures that would be useful to ensure meaningful human control—because, as I said at the outset, I think there is much common ground between the position that we have articulated and the position that has been articulated by those who say we need a norm of meaningful human control.

3 *Supra* fn. 1.

But, I think there are some key reasons why the U.S. view is that "meaningful human control" simply is not an adequate way to describe what is needed for responsible use and development of LAWS, and we continue to think that the term "human control" risks obscuring some of the genuine challenges that these technologies present.

First, no one can really agree on what "human control" means. In discussing this issue at the GGE, there have been almost as many different ways of describing "human control" as there have been delegations in the room. So it has not proven to be a useful construct for building consensus among members of the GGE.

Second, it is also not a very useful umbrella term as a practical matter. How a weapon system is controlled is often very specific to the particularities of that weapon system, and control systems can vary greatly from system to system. This is part of the reason why past regulation of weapons systems under IHL has not included broadly applicable standards for weapon control—the concept in practice does not work very well across different types of weapons.

Third, I believe the concept of human control mistakes the "means" for the "ends." Existing IHL instruments, such as the CCW and its Protocols, do not seek to enhance "human control" as such. Instead, they seek to ensure that the use of those weapons is consistent with IHL. Although control over a weapon system can be a useful way in certain circumstances to ensure that a weapon system is used in compliance with IHL, it is not the only way or always the best way to do so; that is, "control" is not, and should not be, a means in and of itself—but rather one of many ways that states can consider in best effectuating human intent in the use of a weapon system.

Some may say that it is important to emphasize "human control" because they view developments in the use of automation

or autonomy in weapons system as actually decreasing human control over the use of force. And I think we would say that this emphasis is not necessarily correct.

Technical sophistication does not necessarily mean that there is any less human involvement in the decision-making regarding how a weapon is used. In fact, the whole point of some of this technology, for example, sensors and computers, is that it allows humans to have more options for when, where, and how force is used; that is, essentially to make judgments about using force and to have the operator's intent and judgments effectuated by machines without the operator being required to control every step of that machine's process manually. Automated or software control systems can also reduce the degree to which effectiveness in the execution of those important decisions depends on the perception and skill of an operator, which can be negatively impacted in combat by various factors, such as fatigue, fear, or deception. And the use of "smart" weaponry with autonomous functions has actually, I believe, in many ways increased the degree of control that states exercise over the use of force. For example, by increasing the precision of the execution of decision-making, the operator arguably is ensuring better control over the use of force, even though it is not through manual control of every step of a weapon's deployment and use. Theoretically, another way to think about it is that, if an operator might be able to exercise control over every aspect of a weapon system, but the operator is only reflexively pushing a button that is recommended to him or her by the system, the human is not really exercising any judgment, even though the human operator is exercising control in pushing a particular button. What we are looking for here is human intention and human judgment, not necessarily control.

On the other hand, judgment can be implemented through the use of automation. For example, use of algorithms or even autonomous functions that take control away from human operators

can better effectuate human intention and avoid accidents. One system that is a useful case study is the Automatic Ground Collision Avoidance System, which was developed by the U.S. Air Force in order to help prevent "controlled flight into terrain" accidents. The system essentially assumes control of the aircraft when an imminent collision with the ground is detected and then returns control back to the human pilot once the collision is averted. This can help avoid accidents through an automatic feature that actually removes control by the human operator briefly in certain circumstances. Another example would be certain defensive autonomous weapon systems, such as the AEGIS Weapon System and Patriot Air and Missile Defense System, which have autonomous functions that assist in targeting incoming missiles. The machine can strike incoming projectiles with much greater speed or accuracy than a human gunner could achieve manually—so, although the human may not manually control the speed at which the machine is operating, the human is still exercising judgment over the use of force. The machine is really just effectuating that intention and that judgment more efficiently than a human could do so himself or herself.

Finally, some might argue that it is important to emphasize control because of concerns that the use of autonomous weapons systems somehow removes individuals from responsibility for decisions to use force, which are some of the gravest and most serious decisions that a human being can make. But we do not believe this is true. Human actors are responsible for their decisions to use force regardless of the nature of weapon used. The lack of manual control over a weapon system does not remove this responsibility or create an accountability gap. This is, in fact, recognized in the GGE's second Guiding Principle.[4] Machines may be able to synthesize data and

4 Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Sept. 25, 2019, U.N. Doc. CCW/GGE.1/2019/3, Annex

apply algorithms faster than a person could, and they may be able to do so more accurately. But machines are not moral agents, and human beings do not escape responsibility for their decisions by using a weapon with autonomous functions to execute those decisions, in the same way that human beings do not escape responsibility for taking a life with a knife or a gun rather than with their bare human hands.

So, there is no need to stigmatize autonomy as either preventing humans from being held accountable for their decisions, or as inherently reducing control: autonomy does not necessarily do either one.

What next for human-machine interaction?

So, with all of this said, what should come next with regard to human-machine interaction? The reality is that technology is developing rapidly, and standards developed based on our understandings today could be obsolete by tomorrow. So, we need to focus on how to ensure that weapons incorporating those technologies are used in compliance with IHL, and used responsibly, tomorrow. How can we do that?

The U.S. view is that states should take a proactive approach in addressing human-machine interaction. States seeking to develop new uses for autonomy in weapon systems should be affirmatively trying to identify and address these issues in their respective processes for managing the life-cycle of the weapons. One way to do this is to emphasize the importance of weapons review policies and practices—if states are thoroughly and properly conducting reviews of their systems during development and prior to use, they can assess

IV, p. 13, Guiding principle (b) ("Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system;").

whether the specifics of that system can be used consistently with IHL rules and principles, and can be used in a responsible manner.

For example, the DoD Directive requires senior officials to review weapon systems that use autonomy in new ways. This review, which is additional to the normal weapons review processes, is required before a system enters formal development and, again, before fielding, to ensure that military, acquisition, legal, and policy expertise is brought to bear as these new types of weapons are being developed. You have heard me mention reviews several times during the course of my remarks today, but I will say it once more: robust review policies and procedures to ensure lawful and responsible use are one of the most effective ways we can think of to ensure that weapons that are developed tomorrow, and next week, and next year, are used lawfully and responsibly.

Another way to do this is by working to clarify how existing IHL applies to particular systems—the United States developed a paper for the GGE in March 2019 that worked through three general scenarios for the use of autonomous functions in weapon systems and how IHL would apply to those three scenarios.[5] More work could be done on this if states are willing to share their intended use scenarios and their interpretations of how IHL would apply in such cases.

But in our view, there is no better place to do this than in the GGE—and the United States continues to see real value in the conversations that are happening at the GGE, talking through these very difficult issues in a forum that includes technological, military, and legal experts from governments, as well as participation by those from outside of governments. The GGE is really a remarkable

5 U.S. Working Paper, Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems, March 28, 2019, U.N. Doc. CCW/GGE.1/2019/WP.5, available at: <https://undocs.org/en/CCW/GGE.1/2019/WP.5>.

and unique venue: a standing body with a mandate to discuss this extremely complicated, politically fraught topic in a non-politicized way that is grounded in IHL. Where else do we have a body so well-suited to be working through these difficult issues? And, in that light, the United States looks forward in particular to continuing these conversations over the course of the next two years. And we look forward to contributing to a strong outcome before the end of the current two-year mandate of the GGE. Thank you.

Meaningful Human Control over Weapons Systems that Apply Force Based on "Target Profiles"

Elizabeth Minor and Richard Moyes[1]
Article 36 / Campaign to Stop Killer Robots

Introduction

After several years of international discussions on "autonomy" in weapons systems, states appear to have identified that the core area for collective work and agreement is in discussing the aspects of human control (or human-machine interaction) that are necessary during the use of weapons. This has emerged in the discourse as the

1 This chapter, and the presentation by Elizabeth Minor at the Rio Seminar on which it is based, draws from and builds on the following papers by Article 36: "Target profiles as a basis for rule-making in discussions on autonomy in weapons systems," (August 2019), available at <http://www.article36.org/wp-content/uploads/2019/08/Target-profiles.pdf>; "Targeting people," (November 2019), available at <http://www.article36.org/wp-content/uploads/2019/11/targeting-people.pdf>; and "Autonomy in weapons systems: mapping a structure for regulation through specific policy questions," (November 2019), available at <http://www.article36.org/wp-content/uploads/2019/11/regulation-structure.pdf>.

central issue requiring attention, irrespective of whether countries believe new regulation in this area is required or not.

There is still a range of understandings amongst states of what exactly should be included within this work. Some countries have focused narrowly on future systems that could use advanced computer-processing techniques in the application of force. Others have identified wider and more near-term challenges to be addressed, including with potential uses of existing weapons systems. Furthermore, there is some recognition that trying to define a set of specific known or predicted weapons systems to which regulation or principles might apply is unlikely to be a useful exercise, given the very wide range of applications of technology that may raise concerns about adequate control over the use of force now or in the future.

This chapter suggests that, in moving forward, the broad scope of systems underpinning current discussions, which use a particular *process* to apply force—that of matching sensor inputs to a "target profile" of characteristics, without further human action or intervention within a particular "envelope" of space and time—gives a useful starting point for focusing on what the key concerns are, and how these might be effectively addressed through agreement on principles that can stand the test of time and future developments. This scope covers all the systems and possibilities about which concerns have been raised during the international debate on "autonomy" in weapons systems so far. It also gives the building blocks for considering and elaborating the key aspects of sufficient control or interaction. Addressing questions such as those over temporal and spatial limits in the deployment of weapons systems, and the need for an adequate understanding of how systems will apply force and the contexts in which they operate, can be approached from this starting point.

In this chapter, the authors firstly situate discussion in the range of concerns that have been raised about increasing "autonomy" in weapons systems, and the need to adequately respond to them. The chapter then suggests that the broad scope of systems that use "target profiles" to apply force can be a basis and tool for considering further work and regulation. The authors then look at the key challenge with all these systems for a human commander—uncertainty over where, when and to what force will be applied—and the areas and questions that therefore merit attention in order to create structures for regulation that can ensure meaningful human control is maintained over such systems.

The range of concerns expressed about "autonomy" in weapons systems

International discussion on moving towards regulation or other agreement in the area of "autonomous" weapons systems should be grounded within the range of concerns that have been expressed in the international debate about "autonomy," including increasing automation in the application of force and the reduction of human involvement. The range of concerns expressed by states, international organisations, scientists, philosophers, and other civil society provides the basis for international engagement on this issue, the reasons for states to undertake further work in this area, and the aspects that a comprehensive and coherent international response should aim to address.

A review of some of the recent policy-relevant literature and interventions in international forums indicates that concerns can be broadly grouped into those regarding: dehumanisation in the use of force; dangers to civilians; legal challenges; technological concerns; and risks to international peace and security. The content of these concerns can be summarised as follows:

On dehumanisation, many, including Article 36, have argued that employing systems that use sensors to indicate the presence of a person and activate force automatically on that basis involves treating people as objects, violating human dignity.[2] Additionally, concerns have been raised that individuals could be killed by advanced anti-personnel systems directly based on indicators of their gender, race, religion, or other characteristics, which would be unacceptable in additional ways. If sophisticated anti-personnel systems were to use advanced computational processes and training data to determine who triggers activation of force, it has been raised that the gendered, racial, and other biases in their input data as well as their design would, furthermore, almost certainly produce discriminatory harms.[3]

On the dangers to civilians, in a global context where civilians already make up the majority of the victims of war, one concern expressed in this space is that increasing remoteness and automation in warfare further shift the burden of the impacts of armed conflict from the warring parties who employ these technologies onto the affected communities that would continue to suffer the "collateral damage" where war is fought.[4] The use of sensor-based anti-personnel systems would also risk violence to people incorrectly sensed as targets. Their use could risk eroding civilian protection norms more generally, too—for example, if conflict parties shift the onus onto civilians not to enter areas where these systems operate, undermining a presumption of civilian status.[5] Others have argued that increasing automation could, furthermore, marginalise the role of emotion,

2 Article 36 (November 2019), "Targeting people," available at <http://www.article36.org/wp-content/uploads/2019/11/targeting-peopla.pdf>.

3 See for example Hayley Ramsay-Jones (2020), "Intersectionality and racism," in Campaign to Stop Killer Robots Campaigner's Kit, available at <https://www.stopkillerrobots.org/wp-content/uploads/2020/02/2020_Campaigners-Kit_FINAL.pdf>.

4 See for example PAX (2019), "Killer Robots: What are they and what are the concerns?," available at <https://www.paxforpeace.nl/media/files/pax-booklet-killer-robots-what-are-they-and-what-are-the-concerns.pdf>.

5 See Article 36, "Targeting people".

compassion and the ability for individuals to challenge illegal orders in conflict, entrenching violent masculinities and increasing the risks of violence to civilians.[6]

On the legal side, there is consensus among states that humans alone are responsible for applying the law in the use of weapons, and that this "cannot be transferred to machines."[7] Nevertheless, human commanders require control over and understanding of systems and the context of their use in order to make meaningful legal judgments.[8] Uncertainty over where, when and how force will be applied by increasingly "autonomous" systems challenges this. Additionally, if people become legally responsible for complex systems of which they cannot know the full range of effects—for example, because "machine learning" processes that cannot be traced by human programmers have been used to set what the system will apply force to—it will be difficult to hold people accountable in any meaningful way for the results of using these systems. This would create liability challenges.[9]

Technologically, a major concern is that more complex systems become less explainable, predictable, and reliable. This might be a particular concern in the case of systems that are tasked to develop further functions after their initial use, or where the computational

6 See for example, Women's International League for Peace and Freedom (2019), "A WILPF Guide to Killer Robots," available at <http://www.reachingcriticalwill.org/resources/publications-and-research/publications/13601-a-wilpf-guide-to-killer-robots>.

7 Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (2019), "Revised draft final report, Annex III, Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems," available at <https://www.unog.ch/80256EDD006B8954/(httpAssets)/815F8EE33B64DADDC12584B7004CF3A4/$file/CCW+MSP+2019+CRP.2+Rev+1.pdf>. See guiding principles (b). The possible technological capabilities of any system to apply rules are irrelevant to this question, since the law as it is written requires responsibility and accountability from people.

8 See for example remarks by the International Committee of the Red Cross in this volume.

9 See for example Human Rights Watch (2015), "Mind the Gap: the Lack of Accountability for Killer Robots," available at <https://www.hrw.org/report/2015/04/09/mind-gap/ lack-accountability-killer-robots>.

processes used are opaque, for example. Such complexity poses risks of unintended consequences as well as raising moral and legal concerns. The interaction between complex systems may also be highly unpredictable, with potentially dangerous results. Systems operating at speeds far beyond human cognition could also produce unintended outcomes before they can be brought under control.[10] Risks also arise through placing excessive trust in technologies—for example, through automation bias[11]—or if faith is placed in systems to perform functions or to solve human problems that by their nature they cannot—for example, to make human warfare "clean" of human imperfections.

For international peace and security, it has been highlighted in the international debate that ongoing developments in "autonomous" weapons risk the continuation of a new arms race and proliferation amongst states and potentially others. Remoteness and automation could also risk lowered political thresholds for the use of force, due to the lower physical risks to the attacking party.[12] Additionally, competing understandings of what the use of advanced technologies and the implementation of legal principles in relation to them could lead to escalations between states,[13] as could the high-speed responses of systems lacking sufficient human judgement or input.[14]

10 See for example Noel Sharkey (2020), "Fully Autonomous Weapons Pose Unique Dangers to Humankind," Scientific American, available at <https://www.scientificamerican.com/article/fully-autonomous-weapons-pose-unique-dangers-to-humankind/>.

11 See for example International Committee of the Red Cross (2019), "Autonomy, artificial intelligence and robotics: Technical aspects of human control," available at <https://www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control>.

12 See for example PAX (2019) above note 4.

13 For example, UNIDIR has explored this theme in relation to current and next generation UAVs. George Woodhams and John Borrie (2018), "Armed UAVs in conflict escalation and inter-State crisis," available at <https://www.unidir.org/files/publications/pdfs/armed-uav-in-conflict-escalation-and-inter-state-crisis-en-747.pdf>.

14 See for example Noel Sharkey (2020) above note 10.

There are two core issues for the international community to respond to that arise from this broad range of concerns expressed in the international debate:

Firstly, it is necessary to determine whether some of the ways of applying force under the scope of discussion are fundamentally unacceptable—irrespective of whether these systems might be effectively controlled. Article 36 has argued that anti-personnel systems in this area should be prohibited as they present an insurmountable affront to human dignity, as well as unacceptable risks for the protection of civilians.[15]

The second core issue to address is how sufficiently meaningful control over the use of weapons systems can be maintained, in order to: uphold ethical and existing legal principles; prevent increased risks to human and state security; and minimise harm to individuals. Additionally to upholding the law, which has seen great emphasis in international discussion, keeping this control has clear moral dimensions: it entails making decisions about what the limits of acceptable behaviour are where force is to be applied based on the set process of a machine to objects and phenomena, including people. As a focus, Article 36 and others have argued that maintaining meaningful human control over individual attacks should provide the central point of discussion on human/machine interaction, in order to concentrate on the key site of decision-making and legal responsibility. Control in the design of systems, as well as in their use,[16] are both relevant to control over individual attacks.

15 Article 36 (2019), "Targeting people," available at <http://www.article36.org/wp-content/uploads/2019/11/targeting-people.pdf>.

16 IPRAW has set out a conceptual approach to control in design and use in systems—see for example IPRAW (2019), "Focus on human control," available at <https://www.ipraw.org/wp-content/uploads/2019/08/2019-08-09_iPRAW_HumanControl.pdf>.

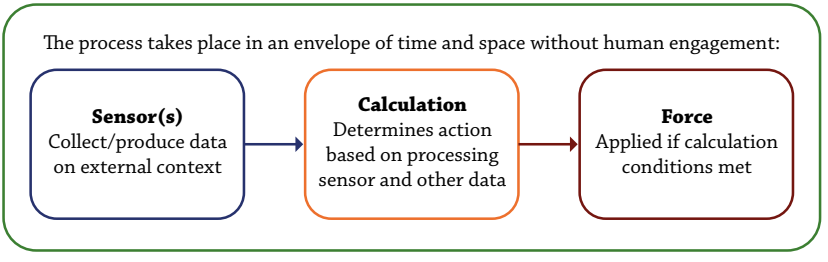
A scope of systems that use "target profiles" underpins current discussion

Systems that apply force through a process of matching sensor inputs to a "target profile," without human action or intervention, within a certain envelope of time and space, are the broad scope that underpins current international discussions. All concerns that states have so far expressed about "autonomy" in weapons systems fall within this area. Taking this as the scope for further discussion gives a building block for defining what should be subject to regulation or the elaboration of principles, whilst avoiding detailed definitional conversations that should take place if and when regulation is negotiated. It gives a clear conceptual focus and basis for forming policies on the rules that will need to be applied. Orienting to this scope also allows a focus on the human role as the boundary issue that countries need to define and negotiate—rather than for instance concentrating on the difference between "automation" and "autonomy" in weapons systems. The latter is an area of discussion that does not necessarily create a productive focus on human control, and may in any case be highly elusive to define technologically.

A "target profile" is the set of conditions that must be met for an application of force to occur. This will be a pattern of sensor data that are proxy indicators of a "target object" or phenomenon, encoded into a system—for example, an object's weight, its heat-shape, or radar signature. Such representations could be more sophisticated in the future, with more advanced technologies—for example using human biometrics. A target profile is an expression and approximation of what a system's designer or user wishes to apply force to. It is a translation of a human concept—for example, a fighting vehicle—into criteria that can be encoded into a machine and detected by its sensors—for example, an object of a pre-defined shape and size that emits infrared energy.

The "sensor-calculation-force" process described above and shown in Figure 1 is a central characteristic of the systems under discussion when "autonomous" weapons are considered in international forums. All conceptualisations of "autonomous weapons" currently fall within this boundary, from hypothetical future systems with so-called "higher-level intent" down to border-based sentry robots that are set to fire when proxy indicators for a person are detected by the system's sensors. Concentrating on systems that use this *process*, rather than focusing on particular types of *weapons* or *technologies*, allows a discussion that encompasses the full range of states' and others' concerns, and could be more resilient to future technological developments and applications. Systems that use a "sensor-calculation-force" process may be land-, sea-, or air-based, and may have different levels of technological sophistication: the potential issues they pose are in how this process is applied and managed—which is central to the concerns about "autonomy" in weapons systems.

Figure 1: The "sensor-calculation-force" process



Situating these systems and concerns about "autonomy"

Systems that use this process of applying force already exist. They include various mines—at the technologically basic end of the spectrum—as well as some of the more sophisticated weapons

in countries' arsenals today, such as missile defence systems and loitering munitions.

Some of these weapons have already caused significant concerns that have led to their prohibition or regulation. Many of these concerns have arisen at least in part from the unintended consequences and harm that can occur when the sensor-calculation-force process—and the automatic application of force—is outside of the effective control of a system's user. For example, various anti-personnel landmines prohibited by the Anti-Personnel Mine Ban Convention will be triggered automatically based only on the weight of people and things passing over them, long after they have been emplaced by the user and the intended targets that match this weight profile—enemy soldiers—are present. This has caused huge indiscriminate harm to communities in areas contaminated by landmines, leading to their prohibition. Other types of mines, and systems using sensor-fuzed sub-munitions, are also subject to regulation under international frameworks.[17]

Systems using a sensor-calculation-force process already raise potential moral and practical challenges for acceptable use and effective control: these challenges are not limited to the most futuristic "autonomous" systems. Discussion about "autonomous" weapons has mostly been focused on possible near-future technologies using a sensor-calculation-force process that might take a step further in the sophistication of their target profiles and how these are constructed, or systems that could operate with less human intervention due to advances in technology. The main concerns are that such future

17 See Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction (1997), Protocol on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices as amended on 3 May 1996, Protocol to the Convention on Certain Conventional Weapons (1996), Convention on Cluster Munitions (2008), available at <http://disarmament.un.org/treaties/>.

systems could intensify existing or pose new ethical, moral, legal, and protection hazards.

The challenges that flow from uncertainties, and the building blocks for a response

The challenges of control with systems that apply force through matching sensor inputs to a target profile within an envelope of time and space where there is no human evaluation of the sensor data all flow from uncertainty. By the nature of such systems, a human commander will not know *exactly* when, where or to what or whom force will be applied when the system has been switched on, emplaced, or released—because the force application itself is activated by the matching of an encoded profile with sensor input data absent further human intervention. If the period of time during which force could be activated is longer, and the opportunity for interaction between the system and phenomena in its environment is greater, this uncertainty will increase. This makes the possibility for unintended outcomes more likely, as well as increasing the potential for ethical challenges.

Understanding the different components of this uncertainty gives a starting point for describing the questions and building blocks required for a human commander to maintain meaningful human control over weapons systems that apply force using target profiles, in individual attacks.

Who or what might trigger force?

What or whom force will be applied to by a sensor-based system depends on what will actually trigger force in practice. Things that activate force will include both the targets a user intends the system to strike, and unintended objects or phenomena that fall within a system's target profile, based on the pattern of sensor data that this contains. The uncertainty here represents the gap between human

intent and system performance. As no system can replicate human intent—this being a meaningless proposition from a technological perspective—understanding the relationship between a system's encoded target profiles, the "target objects" it aims to apply force to, and the unintended "remainder objects" that will also trigger a force application, is vital to understanding the likely outcomes from the use of a specific system—and so to making meaningful judgements about if, where and how it can be used.

To give some illustrative examples of the intended and unintended objects that fall within the target profiles of existing systems, the encoded target profile of an anti-vehicle mine may represent the human intent to strike a "military vehicle," represented as a weight, for example, over 150kg. Military vehicles over this weight will activate the mine, as will many other vehicles—these are the "remainder objects". A ship-borne missile defence system will encode the concept of "an inbound missile" through indicators of radar signature, speed and direction—remainder objects may include other aircraft, as accidents involving these kinds of systems have shown. Sensor-fuzed weapons systems may encode a "military vehicle" as a heat/shape pattern—some other heat/shape patterns may also trigger force, such as other vehicles. For many weapons systems, the exact parameters of a system's target profiles will be commercially or militarily sensitive—which makes it harder for the efficacy and other aspects of a system to be evaluated and assessed.

If the range of unintended remainder objects that will activate force is large, this may be one indicator of concern relating to sufficient control—and a potential challenge to legal assessments such as decisions on proportionality. Furthermore, the profile of a system's remainder objects may render it unacceptable in other ways. For example, if an anti-personnel system that applies force based on human biometric indicators is known to have an error rate of 10% when calculating matches, this means that it will, with

certainty, kill the wrong people 10% of the time. This certainty of error is qualitatively different to unintended "collateral" harm, and might be considered morally unacceptable—even if it might be seen as legally proportionate.

Understanding how a target profile has been constructed is also crucial to understanding what force might be applied to, and for managing uncertainty. Human users of systems need to understand the practical implications of the technical processes systems use: if these processes are opaque, such that sufficient knowledge of the implications cannot be obtained, this poses potentially unacceptable uncertainties and difficulties for sufficient control. For example, if "machine learning" processes are used as a basis for a system's "object recognition" calculations, it may not be possible for a person to understand or interrogate the specific characteristics or features of the objects that will activate force. Experiments involving these advanced computational techniques for developing "object recognition" have shown that "false positives" are often produced that appear unrelated to the task set by human programmers—but by the nature of these systems, the reasons for why the system has performed in this way cannot be traced. A weapons system that is tasked to develop or refine the parameters of the target profile after being emplaced, switched on or released would also generate further uncertainties for human users about exactly what it would apply force to.

These various aspects of uncertainty raise questions about how objects and phenomena can reasonably be encoded as targets—and by what or whom. This is a practical and technical question relating to the gap between human intent and system performance, and the volume of "remainder objects" a system may strike. It is also related to the possibility of generating further uncertainties through using increasingly complex systems to construct target profiles, marginalising the human role in this aspect of system design.

Furthermore, it is a legal, moral, and ethical question of what can acceptably be treated as an object in the use of a weapons system.

When sensor-based systems that use target profiles to apply force are used, there will also be uncertainty about which other things or people in its surroundings might be affected when force is triggered by a target or remainder object. The number of applications of force a system will make in one human-mandated "attack"—for example, the number of munitions or projectiles a system will release—and the destructive power of these applications also increases uncertainty about the exact effects that will result from a sensor-based system—whether these are intended or not, and to targets or surrounding objects and people. These uncertainties are linked to those about when and where exactly force will be activated.

When and where might force be applied?

Uncertainty about exactly where and when force might be applied in the use of sensor-based systems raises questions about how control in time and space can be ensured that is sufficient for upholding ethical principles and making meaningful legal judgements. In international discussions at the Convention on Certain Conventional Weapons on "autonomous weapons" as well as elsewhere, and in military practice, it has already been widely accepted that weapons systems should be used in a physical space and duration of time that is in some way set by a human commander.

Uncertainties in this area relate to factors in the context of a system's use, and how the system will interact with this context. Unintended consequences resulting from the force applications of sensor-based systems will generally become more likely the wider the area over which a system is used, and the more "complex" the environment—populated areas, where there are high numbers of people and their infrastructure, represent one type of more complex environment. This is due to the range of objects that might be

present and affected by a system in wider or more complex areas. The duration in time over which the sensor-analysis-force process is in use also contributes to uncertainty, and interacts with this. A longer duration generates greater uncertainty about what might be affected by the system, as different objects and people move in and out of a particular area over time, for example.

Additionally, the further in time an actual application of force occurs from a person making a legal judgement about the use of a system, the less meaningful this judgement will become. The information about the context of a system's use that a decision-maker will have relied on will have become increasingly out-dated.

Uncertainty—and so the risk of unintended harmful consequences—in the use of sensor-based weapons systems becomes greater, in summary: the more out-dated and less relevant the contextual information and assessment under which a system has been released is; the wider a system's area of use; the longer the duration for which a force application by the system is possible; the greater the complexity of the context in which the system is being used; the greater the number of applications of force a system will undertake in one use, or attack; the larger the destructive power of the system; the greater the number of "remainder objects" that may fall within a system's target profile; and the less the practical implications of its target profile are meaningfully understood.

To give an example illustrating some of these issues, consider a sensor-based weapons system that is used to strike the general target of a group of fighting vehicles. The "target object" for the system in this case is one fighting vehicle, for which the system searches, matches, and applies force in a defined search area range, for the duration that it can travel after release. If such a system can strike two or three objects during one use, the question arises, if the fighting vehicles become more dispersed, at what point it becomes

unreasonable to consider this group of vehicles as a single target of one attack. This question is intensified if the group of vehicles become more widely dispersed, and across a populated area, where there may be specific collateral damage concerns for protected objects, more civilian vehicles that might fall under a system's target profile, and a generally diminished ability for a commander to evaluate likely civilian harm from the system's force applications. Uncertainty and the possibility of unintended consequences are increased where the concept of a single attack using a system is stretched in time, space and the context for force that can or has been evaluated.

Key questions for a human commander for retaining meaningful human control

For meaningful human control and meaningful legal judgements, this discussion implies that principles on the spatial area and duration of use for sensor-based systems must be set such that a commander can fulfil their legal obligations in relation to an attack; that their legal judgement must be sufficiently proximate to an application of force to be relevant; and that they must have specific understandings of the intended and unintended effects of the systems used. From a legal and ethical perspective, the core aspects a human commander is required to understand are how a system operates, and the context of its use.

Some key questions for a human commander using sensor-based system are therefore:

- What intended and unintended objects and phenomena fall within the system's target profile?
- Howis the target profile constructed, and is this understood?
- How can space, duration, and time of operation be controlled effectively and meaningfully, including to ensure sufficient contextual knowledge for legal judgments on a specific attack?

- What "quantity" of force will the system apply (e.g., how many munitions will be release separately or onto separate objects)?
- What are the risks to civilians (which should be verified by the commander)?

Ordered another way, some key elements[18] for conceptualising meaningful human control arising from this discussion are the need for:

- Predictable, reliable, and transparent technology;
- Accurate information for the user on the outcome sought, the technology, and the context of use;
- Timely human judgment and action, and a potential for timely intervention; and
- Accountability to a certain standard.

In general, a key question in this area that should produce decision points is whether human analysis, judgment and oversight are retained, or reduced, with the introduction of a new technological capability or system. The discussion and elements above may be considered implicit in the principles of existing international law, or represent how many states might interpret existing law. However, it is clear from international discussion so far that there is not consensus on what constitutes an effective approach to human control, which is why states have identified it as the key area for further discussion. Additionally, existing law does not currently have all the answers to the concerns raised about "autonomous" weapons—including whether certain specific developments and applications in this area are of central importance and morally unacceptable, and, furthermore, must be prohibited altogether.

18 Discussed in Heath Roff and Richard Moyes (2016), "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons," available at <http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf>.

Conclusion

Moving forward, focusing on a scope for discussion of systems that apply force through a process of matching sensor inputs to a "target profile" without human action or intervention during a particular envelope of space and time allows a focus on what the key elements are that make control sufficient in weapons systems, and what limits to control would make a system cross the line of acceptability. Considering the key factors producing uncertainty and the risks of unintended consequences, and the key questions for human commanders that these generate, gives the building blocks for considering elements of meaningful human control over systems that apply force based on "target profiles."

In approaching regulation, a structure will be needed within this broad scope of systems that engages with the fact that there is not a simple set of technologies that can be prohibited or regulated as full, complete systems in this area. There is, rather, a range of applications, usages, and assemblages that in different combinations or deployments could cross lines of acceptability or not. Article 36 has suggested[19] that constructing an effective regulatory structure would likely involve taking a broad scope of systems for regulation—those using the sensor-calculation-force process—and applying within this scope prohibitions on certain system configurations, as well as obligations on other system configurations regarding their development and use to ensure that these can be used in accordance with established legal obligations and principles.

Systems that use sensors to apply force against human beings, as well as systems that are complex in their functioning to the point that they cannot meaningfully be controlled, are two examples of

19 Article 36 (2019), "Autonomy in weapons systems: mapping a structure for regulation through specific policy questions," available at <http://www.article36.org/wp-content/uploads/2019/11/regulation-structure.pdf>.

configurations that could be prohibited. Such prohibitions would help address many of the concerns over dehumanisation in the use of force, technological and legal challenges that have been raised in the international debate. Other systems within this scope should be subject to obligations that ensure meaningful human control is maintained over them. These obligations would help to address many of the concerns around dangers to civilians and legal issues raised by "autonomy" in weapons systems. The fact of a regulatory structure itself would go a considerable way towards addressing the international peace and security concerns raised by increasing "autonomy" in weapons systems.

Adopting such a structure for regulatory approach would help states to address the problem of increasing "autonomy" in weapons systems comprehensively, in a way that is technologically agnostic and future-proof, and that would address the core concerns raised by the range of stakeholders in the international policy debate. As countries go forward in their discussions about human control or "human machine interaction" in the context of weapons systems, they must consider how principles can be translated into the elements of a meaningful regulatory structure, and begin to make concrete proposals on these elements, in order to construct an effective response.

Meaningful human control over weapons systems that apply force based on ‘target profiles’

Elizabeth Minor, Advisor, Article 36
Rio Seminar on Autonomous Weapons Systems
February 2020

Article36

Different concerns expressed about 'autonomous weapons':

1. **Dehumanisation**
   - Treating people as objects; Bias; Gender, race, other discrimination

2. **Danger to civilians**
   - Displacement of violence from militaries onto civilians; Erosion of civilian protection norms; Marginalisation of compassion

3. **Legal challenges**
   - Uncertainty about when/where/how systems will apply force, as time/area of activation increases
   - Need for human control, understanding of systems and context to make legal judgments
   - Liability issues

4. **Technological concerns**
   - Explainability, predictability, reliability decreases with complexity
   - Danger from high speed systems, and interaction between complex systems
   - Danger of misplaced trust and faith in technology

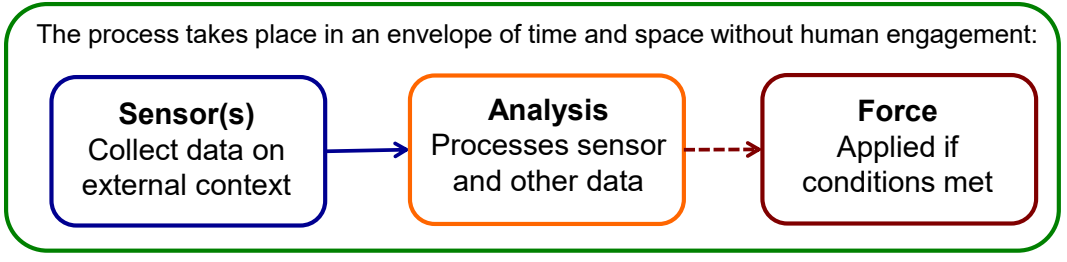
5. **Risks to peace and security**
   - Arms race and proliferation; Lowered political thresholds for the use of force; Escalations and crises from competing legal understandings and high speed reactions of systems

Core issues to respond to in the area of ‘autonomous weapons’:

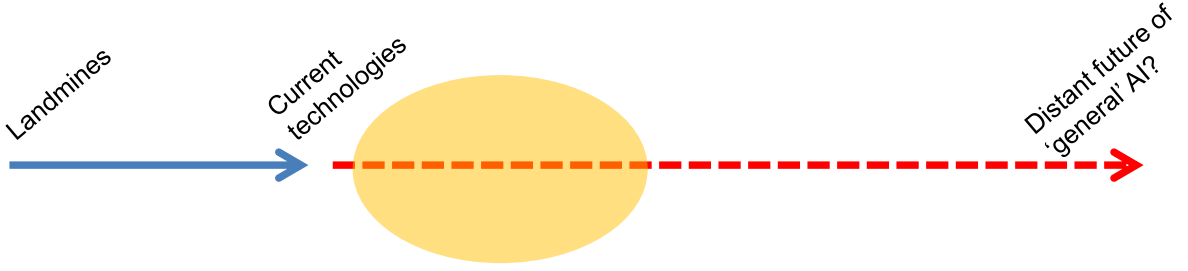
- Are some of these ways of applying force fundamentally unacceptable?
- How can we keep sufficient control over the use of weapons systems?
  - Ensuring ‘meaningful human control’ over attacks

Broad scope underpins discussions:

- Systems that apply force through a process of matching sensor inputs to a 'target profile,' without human action/intervention:



Situating these systems and concern



Phalanx (US)

Our challenges flow from uncertainty:

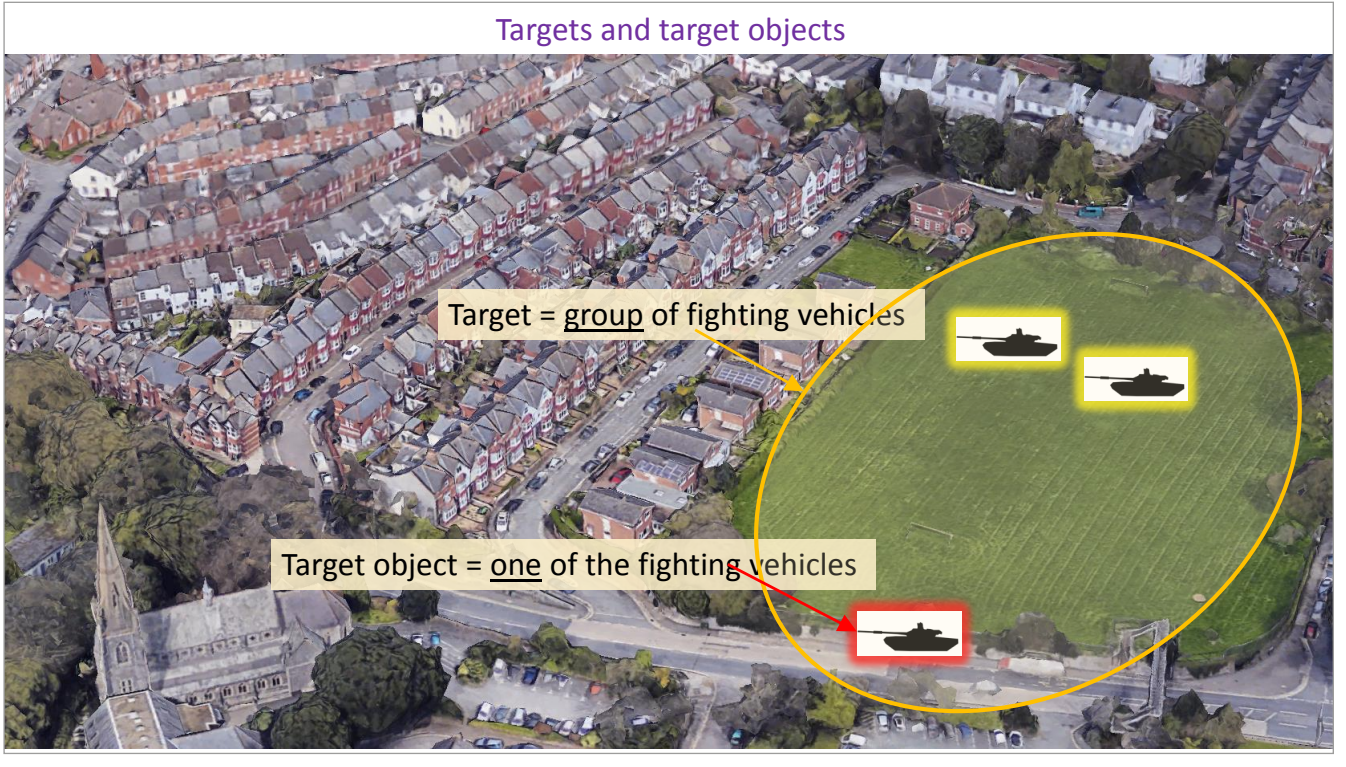
- When, where, to what/who might force occur?
  - What (else) will trigger force in practice?
  - How can things reasonably be encoded as targets?
  - How can we ensure control in time and space sufficient for upholding ethical principles and making meaningful legal judgments?

What/who: target objects, target profiles and remainder objects

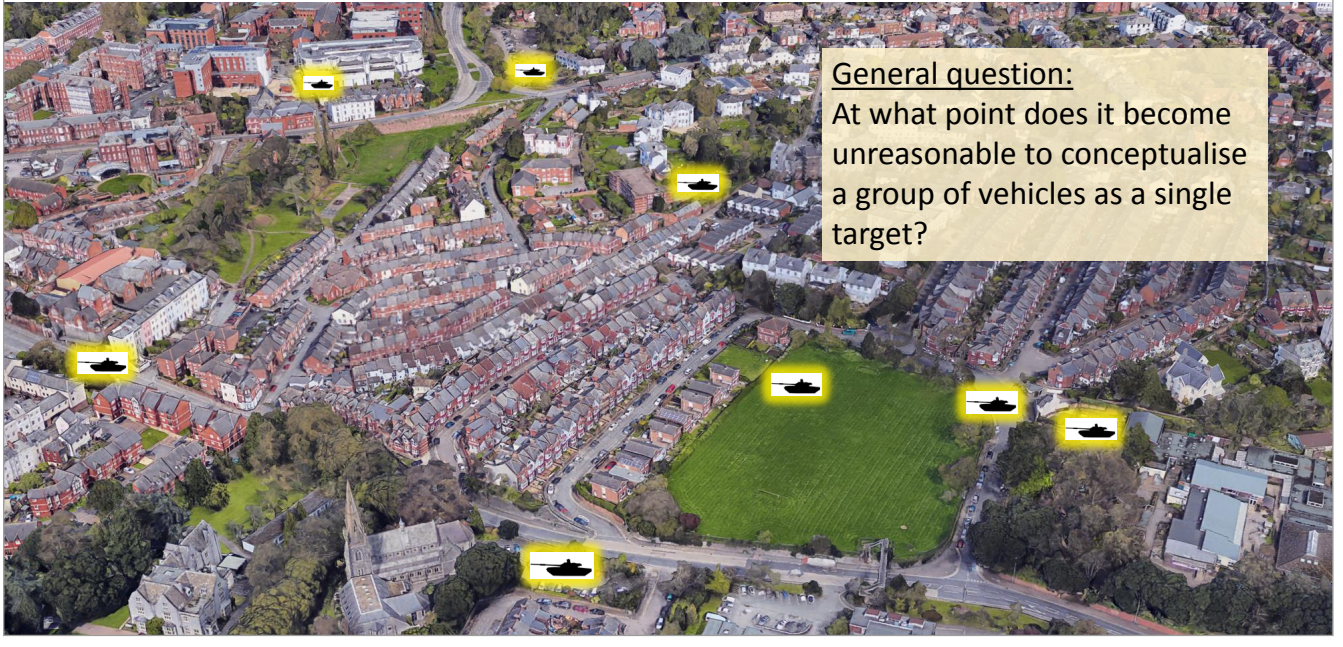
Weapon type	Encoding of target profile(s)	"Remainder objects"
Antipersonnel landmine	A military person encoded as weight > x (approx. 10kg)	All non-military people, many animals, most vehicles
Antivehicle landmine	A military vehicle encoded as weight > x (approx. 150+ kg)	Many other vehicles
Ship-borne missile defence system	An inbound missile encoded as radar signature speed and heading	Some other aircraft (specific parameters not known)
Sensor fuzed weapons systems	A military vehicle encoded as heat/shape pattern	Some other heat patterns, e.g. other vehicles (specific parameters not known)

Where/when?

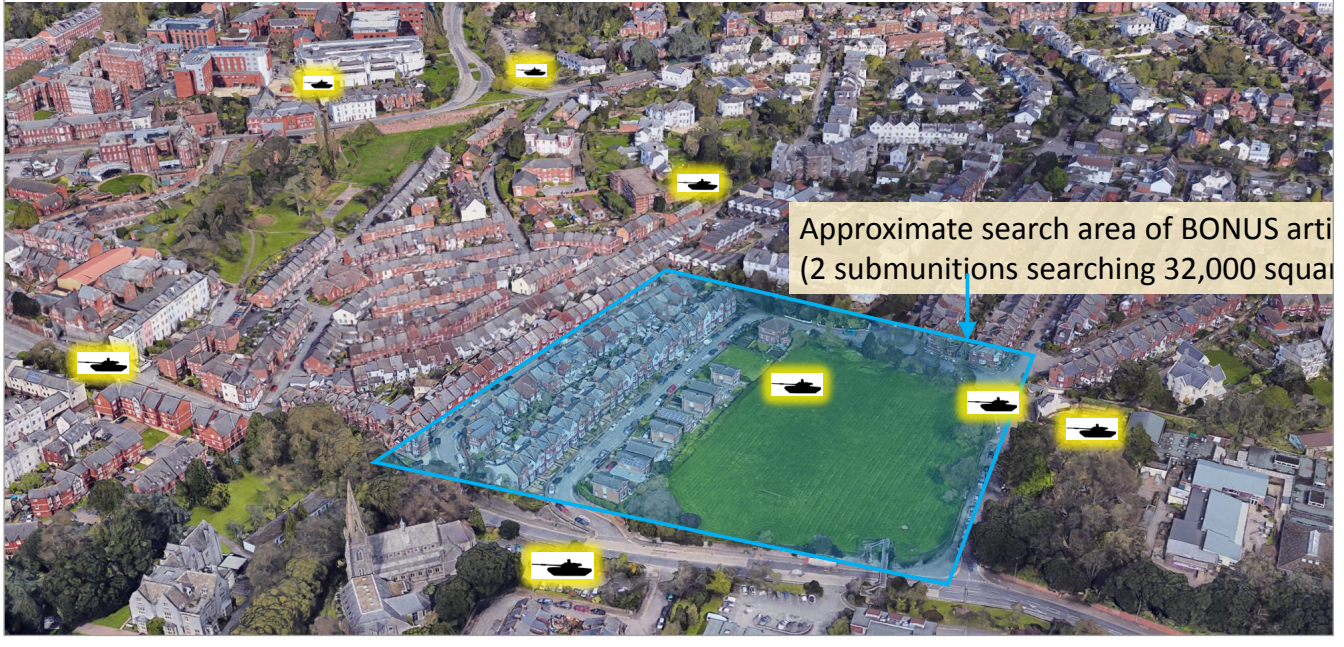
- Physical space – area and complexity
- Duration in time
- Number of applications of force in one 'use'
- Closeness to legal judgment made about use of system for attack


Targets and target objects
Target = group of fighting vehicles
Target object = one of the fighting vehicles

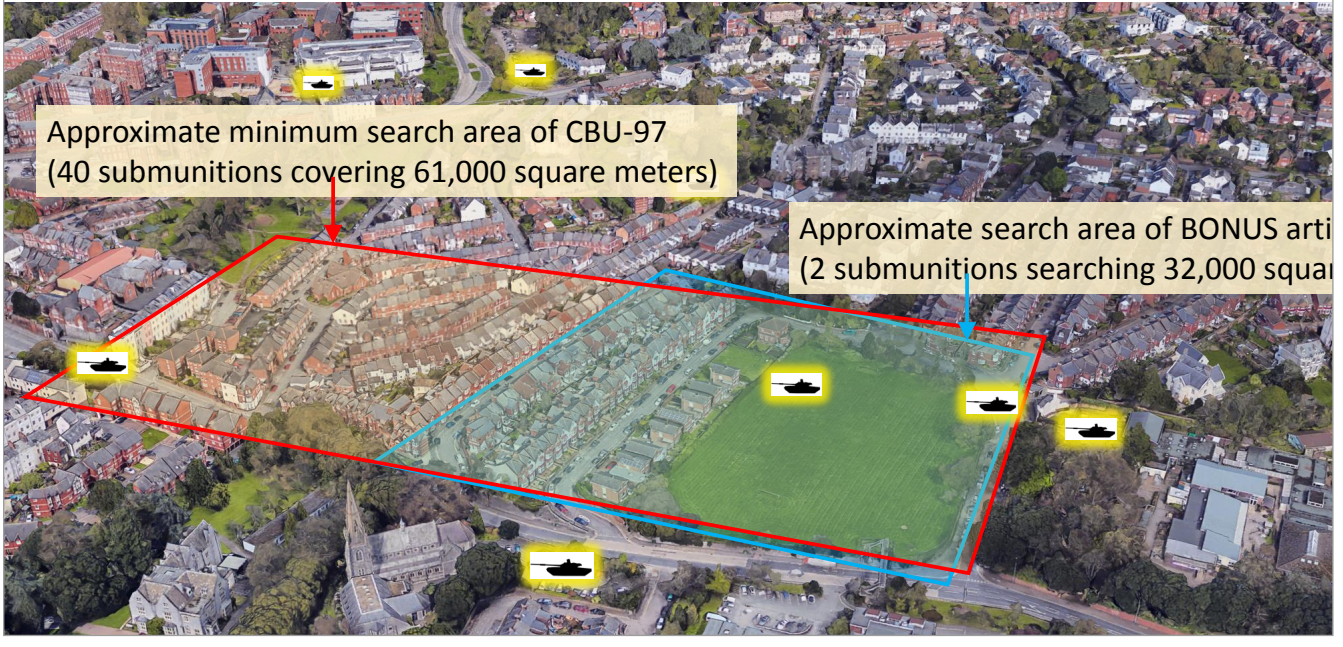
CONCEPTUALISING "TARGETS" / "TARGET OBJECTS" – 2. MORE DISPERSED OBJECTS



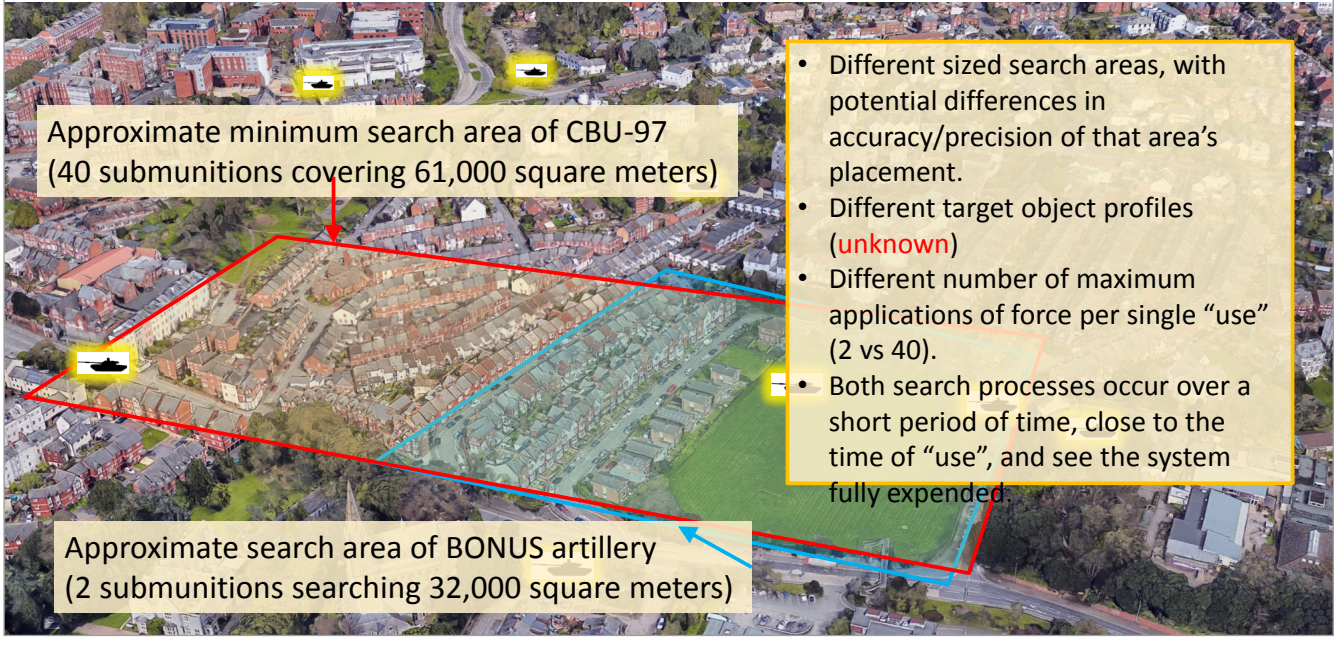
CONCEPTUALISING “TARGETS” / “TARGET OBJECTS” – 2. MORE DISPERSED OBJECTS



CONCEPTUALISING "TARGETS" / "TARGET OBJECTS" – 2. MORE DISPERSED OBJECTS



CONCEPTUALISING “TARGETS” / “TARGET OBJECTS” – 2. MORE DISPERSED OBJECTS

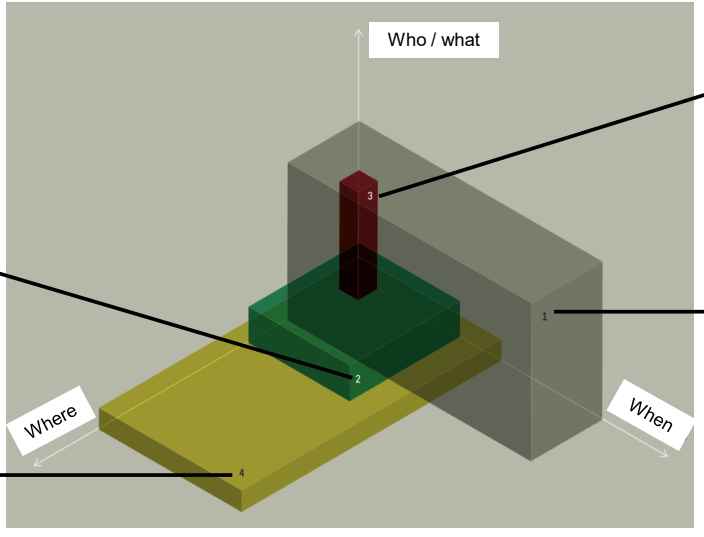


- Different sized search areas, with potential differences in accuracy/precision of that area’s placement.
- Different target object profiles (unknown)
- Different number of maximum applications of force per single “use” (2 vs 40).
- Both search processes occur over a short period of time, close to the time of “use”, and see the system fully expended.

CONCEPTUALISING "TARGETS" / "TARGET OBJECTS" – 3. WIDELY DISPERSED OBJECTS

CONCEPTUALISING "TARGETS" / "TARGET OBJECTS" – 3. WIDELY DISPERSED OBJECTS

Volumes of uncertainty about when, where, who/what



Anti-tank missile with millimeter-wave radar seeker

Anti-missile system

Anti-personnel landmine

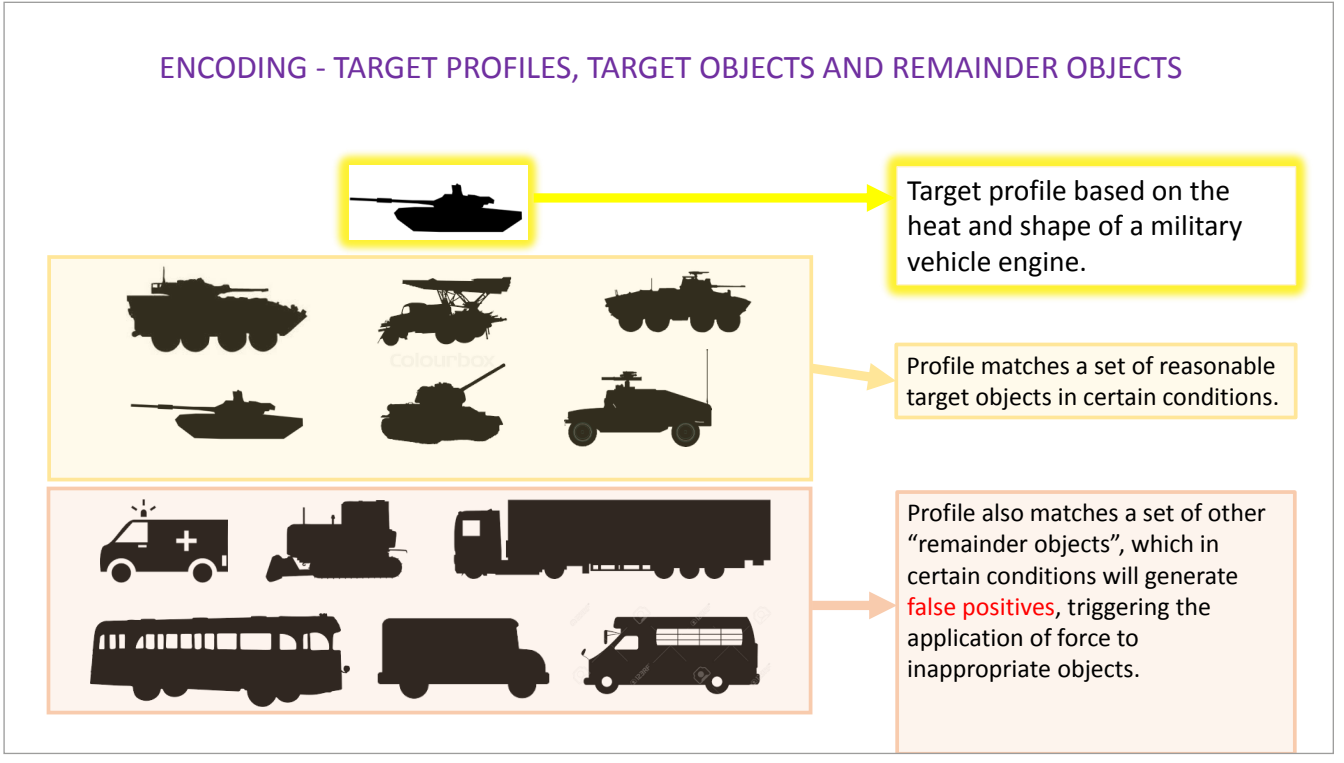
Anti-personnel system using biometrics/ facial recognition

Key issues for human commander:

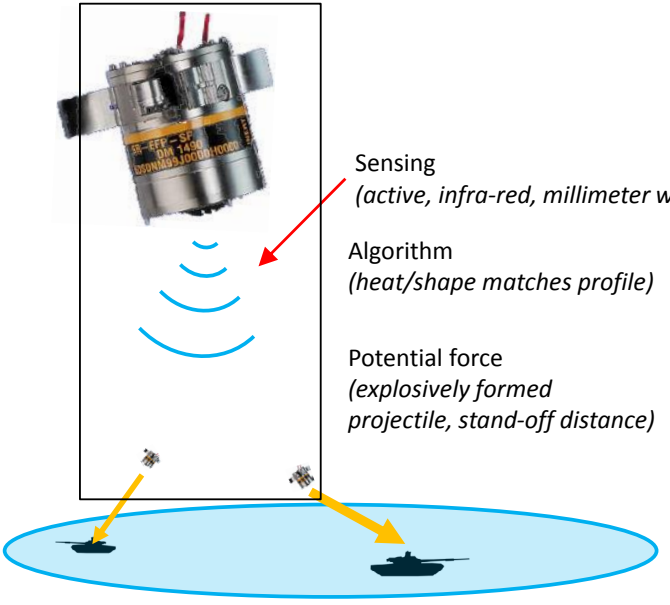
- What falls within the target profile?
  - Intended, unintended and excluded
  - How is it constructed and is this understood?
- Controlling space, duration and time
  - Contextual knowledge for legal judgments on a specific attack
- “Quantity” of force
- Verifying risk to civilians

Key elements of meaningful human control:

- Predictable, reliable and transparent **technology**.
- Accurate **information** for the user on the outcome sought, the technology, and the context of use.
- Timely **human judgment and action**, and a potential for timely intervention.
- **Accountability** to a certain standard.


ENCODING - TARGET PROFILES, TARGET OBJECTS AND REMAINDER OBJECTS
Target profile based on the heat and shape of a military vehicle engine.
Profile matches a set of reasonable target objects in certain conditions.
Profile also matches a set of other "remainder objects", which in certain conditions will generate false positives, triggering the application of force to inappropriate objects.

SENSOR FUZED MUNITION



Sensing
(active, infra-red, millimeter wave)

Algorithm
(heat/shape matches profile)

Potential force
(explosively formed projectile, stand-off distance)

Target profile
a military vehicle encoded as unknown heat/shape profile(s)

Location
Combined search area (approx. 32,000 sq.m) of 2 submunitions, resulting from a separation point that is ballistically produced based on commander coordinates

Duration
Short

Different concerns expressed about 'autonomous weapons':

1. **Dehumanisation**
   - Treating people as objects; Bias; Gender, race, other discrimination
2. **Danger to civilians**
   - Displacement of violence from militaries onto civilians; Erosion of civilian protection norms; Marginalisation of compassion
3. **Legal challenges**
   - Uncertainty about when/where/how systems will apply force, as time/area of activation increases
   - Need for human control, understanding of systems and context to make legal judgments
   - Liability issues
4. **Technological concerns**
   - Explainability, predictability, reliability decreases with complexity
   - Danger from high speed systems, and interaction between complex systems
   - Danger of misplaced trust and faith in technology
5. **Risks to peace and security**
   - Arms race and proliferation; Lowered political thresholds for the use of force; Escalations and crises from competing legal understandings and high speed reactions of systems

Talking Points

Yokoyama Daiki
Conventional Weapons Division – Ministry of Foreign Affairs (Japan)

1- Japan's position on LAWS:

- Japan has no intention to develop a fully lethal autonomous weapons system;
- Meaningful human control should be retained for lethal weapons systems; and
- The exploration of LAWS should be balanced between humanitarian and security perspectives.

2- Japan's position on the human element of LAWS

- Further exploration is necessary to consider the degree and stage of meaningful human control;

- Meaningful human control is a premise in attributing responsibility/accountability; and
- There is a divergence of views regarding the degree and stage of the human element. It is necessary to elaborate on the issue considering the trend of technological progress in the future.

3- Considerations on the human-element part of the GGE Report 2019 (CCW/GGE.1/2019/CRP.1/Rev.2)

- (Paragraph 22.a) Agreement on the importance of the human element;
- (Paragraph 21.a) Human responsibility can be exercised in various ways across the life-cycle of these weapons systems and through human-machine interaction;
- (the GP (c)) Human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon; and
- (Paragraph 22.c) Divergence of views. Human involvement in the development stage may not be sufficient.

4- Considerations on human-machine interaction

The human-element:

There is a perspective that it is necessary to retain a certain level of human supervision. On the other hand, it is also necessary to consider the potential technological development in the future.

We also need to consider peaceful uses of autonomous technologies and their potential merits for emerging technologies to reduce manpower, casualties, and human errors in operation, and to increase precision in attacks.

The machine element:

There is a perspective that imposes on lethal weapons systems a certain level of capability constraints in the operational area/time to retain characteristics such as predictability/reliability for appropriate operation. But, in this case, we should also pay attention to the technological progress that would change the operational environment in the future.

It is important to consider the appropriate integration system in research & development.

(Reference) Measures in private sector and AI (artificial intelligence) guidelines:

- ISO13482 for personal care robots
- AI guidelines of Japan and EC (European Commission)

5. Other challenges for upcoming exploration of the issue related to LAWS

There are key elements of LAWS that should be enhanced to achieve common understandings among stakeholders, such as the definition of human-machine interaction, as well as human control. They are necessary characteristics that make LAWS lawful, regarding the definition of LAWS, and enable human involvement in the lifecycle of weapons systems, etc. We would like to thank all participants for sharing their perspectives in order to reach a common understanding.

Panel 1: Human – machine interaction and human control; from engineering to IHL

Yokoyama Daiki
Ministry of Foreign Affairs, Japan

1. Our position to LAWS
2. Our position to human element
3. Description of human element in GGE report
4. Consideration for Human-machine interaction

(Reference) ISO13482, AI guidelines

5. Other challenges to upcoming GGE on LAWS

Our Position

- Not to develop fully autonomous weapons systems with lethality (directly kill to human beings)
- Importance of Meaningful Human Control and certain implementation of the Guiding Principles
- International rule making with seeking balance between humanitarian consideration and security perspective

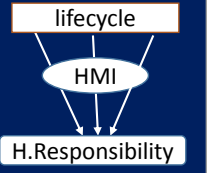
Consideration for human element

- Weapons systems with lethality accompanied with meaningful human control (MHC).
- MHC is a concept that could be used as a premise in attributing responsibility for various effects caused by them.
- There is a wide range of views on where and how much MHC is necessary in the life-cycle of weapons systems.

Consideration through GGE report 2019

- Importance of the human element (Paragraph 22.a)

- Human responsibility can be exercised in various ways across the life-cycle of these weapon systems and through human-machine interaction. (Paragraph 21.a)

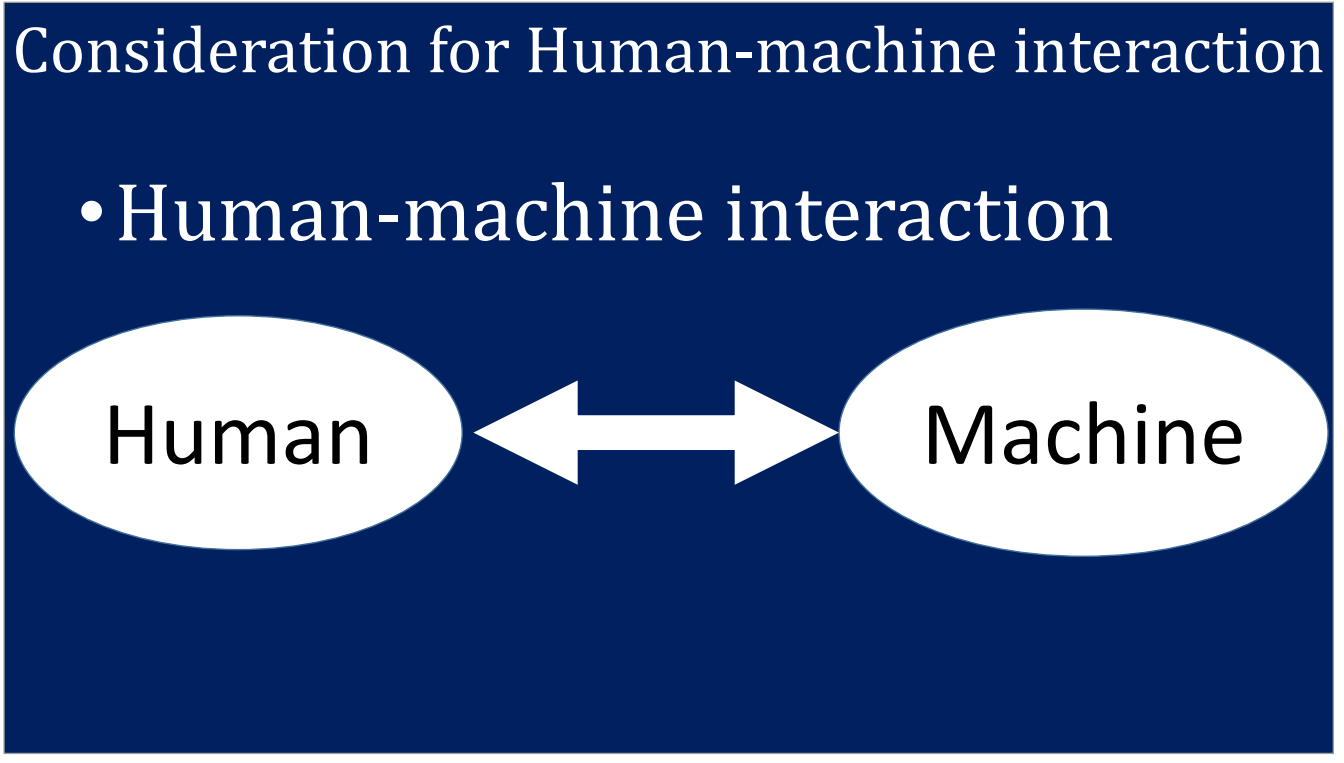


- Human-machine interaction
  various forms at various stages
  should be considered including the operational context, and the characteristics and capabilities of the weapons system as a whole. (the Guiding Principles (c))

Consideration through GGE report 2019

- Human involvement at the development stage may not be sufficient. (Paragraph 22.c)

- Emerging technologies may be useful for enhancing the implementation of IHL (reduce human error and to increase precision in attacks).


Consideration for Human-machine interaction
• Human-machine interaction
Human
Machine

Consideration for Human-machine interaction

- Human element

How Does human have involvement with weapons systems (machine) on their lifecycle?

Political decision, R&D, T&E, V&V, Deployment, Training, Order, Use, Assessment (Example)

Degree and stage of human involvement, depends on technologies and operational context

Consideration for Human-machine interaction

- Human element

Based on the present technological level….

There is a perspective that retaining human supervision of commander in some degree, to ensure chain of command.

Degrees are depends on domains(ground, surface, underwater, aerial) and performance.

It should not be prejudiced the degree for assuring technological development and peaceful uses.

Consideration for Human-machine interaction

- Machine element
- Constrainability in time/area
- Development with appropriate system-integration
- Comprehensive understanding of risks through cooperation between operators and engineers.
- Risk management by engineers through understanding about operational environment and technological maturity (of operators).
- Extensive technological and operational testing and evaluation (prediction for condition of facing imminent risks and effects of countermeasures)
- Tracking evaluation for “lesson-learned” during training, and improve comprehensiveness of risk assessment by reflecting results

Measures in private sector and guidelines

- ISO13482 Safety requirements for personal care robots

Utilizing emerging technologies for social welfare through applying risk assessment and risk reduction process in R&D

KEEP utilizing emerging technologies for maximizing benefits, with minimizing risks with measures

The Guiding Principles for LAWS, and its operationalization, especially (f).

Measures in private sector and guidelines

ISO13482 as reference for engineering of this issue…

KEEP utilizing emerging technologies for maximizing benefits, with minimizing risks with measures

The Guiding Principles for LAWS, and its operationalization, especially (f).

Measures in private sector and guidelines

- AI guidelines

AI Utilization guidelines (Japan)

by The Conference toward AI Network Society, Ministry of Internal Affairs and Communication, Japan

https://www.soumu.go.jp/main_content/000658284.pdf

AI utilization flow (reaffirm the safety in system implementation, risks might occur and even be amplified when AI systems are to be networked.)

Ethics Guidelines for Trustworthy AI (European Commission)

Assessment list

Human Control in the Use of Force

Anja Dahlmann
German Institute for International and Security Affairs (Germany)

The human-machine relation is a crucial element of the CCW debate about LAWS. The following notes are based on the work by the International Panel on the Regulation of Autonomous Weapons (iPRAW), arguing in favor of human control in the critical functions of the targeting process for operational, legal, and ethical reasons. iPRAW defines minimum requirements for human control, in this context, as follows: situational understanding and options for intervention by design and in the use of the weapon system. Those abstract requirements have to be adapted to the specific operational context.

Further work on human control—or, more broadly, the human-machine relation—should be a substantial element of the CCW deliberation on LAWS in 2020/21. This paper highlights a few

aspects that might contribute to the discussion as it focuses on the link between IHL and human control as well as the (legal) notion of the term attack.[1]

Human control as a consequence of IHL: Autonomous functions in weapons systems call for human control before and during the attack. Precaution during attack remains feasible when the operator/commander has sufficient situational understanding and options for intervention along the lines of iPRAW's concept of human control. Accordingly, the operator/commander must be enabled to review legal assessments and translate human decision-making into the system's action during attack prior to the actual engagement.

More precise notion of attack: Defining what constitutes the start of an attack can be useful in unpacking the concept of human control. The most relevant point in the mission thread is not defined by the launch or activation, but by the final necessary decision on target selection and engagement by a human. Weapons systems with autonomous functions potentially move the final human decision to a very early stage of the operation. With regard to the legal judgments to abide by IHL principles, this effect could be challenging for two reasons: first, it can increase the level of abstraction in the target selection process (i.e. class of targets instead of specific target). Second, the environment might change during this extended timespan between targeting decision and engagement, e.g. outdating the initial proportionality assessments.

The underlying notion of attack will, therefore, influence the understanding of the principle of human control in a regulation of autonomous weapons systems. This is because IHL principles like distinction and proportionality are legally required during

1 The following text is drawn from iPRAW (August 2019), Focus on Human Control, available at: <https://www.ipraw.org/human-control/>.

the planning phase, but, to a certain extent, become a question of feasibility in attack. This would alter the need or necessary level of human control in attack.

Context-dependency of human control: While it is possible to develop abstract minimum requirements for human control in the use of force, the appropriate level or implementation of human control depends on the details of the operational context. A "one-size-of-control-fits-all" solution that addresses all concerns raised by the use of autonomous weapons systems will most likely not be achievable because it cannot account for the multitude of combinations of environmental factors, operational requirements, and weapons capabilities. Instead, a (binding or non-binding) regulation would be more useful if it included general approximations to be specified in each case along the lines of existing IHL considerations. iPRAW encourages CCW States Parties to develop and share specific examples for how control by design and control in use can be implemented in weapons systems used in different operational contexts.

Towards Broadening the Perspective on Lethal Autonomous Weapon Systems' Ethics and Regulations

Bianca Ximenes[2], *Diego Salcedo*[3], *and Geber Ramalho*[4]
Federal University of Pernambuco

1. Introduction

LAWS may, without the explicit approval of a human being, decide to cause harm to or kill people. Their adoption involves complex ethical, technical, commercial, legal, regulatory, strategic, and geopolitical issues. That is why, in the scope of IHL, taking into account the CCW, a United Nations GGE has been created to debate the governance of this emerging technology.

2 Informatics Center.
3 Information Science Department.
4 Informatics Center.

The two opposing positions on LAWS could be total banishment or no regulation. Without making any judgment of the value of these two positions, both supported by some countries so far, we prefer to adopt an in-between perspective, since intermediary solutions raise more interesting and complex debate than adopting one of the two positions. Indeed, supposing that these kinds of weapons could be authorized or adopted, some questions should be answered: in which cases can they be adopted? Under which circumstances can they be used? Which kinds of weapons can be fully automatized? How does one limit the damage of the use of such weapons? Who is accountable for their use? What are the adoption criteria and processes for these weapons?

Several advances have been made on LAWS governance by the GGE/LAWS, establishing the basis and premises that may enable an agreement or convention on the topic. In particular, the CCW/GGE has, in their late 2019 session, converged to form the 11 Guiding Principles for LAWS, representing an excellent starting point for more detailed discussion (CCW/GGE.1/2019/3).

Our reflections on LAWS issues are the result of the work of our research group on AI and ethics at the Informatics Center in partnership with the Information Science Department, both from the Federal University of Pernambuco, Brazil. In particular, our propositions and provocations are tied to Bianca Ximenes's ongoing doctoral thesis, advised by Prof. Geber Ramalho, from the area of computer science, and co-advised by Prof. Diego Salcedo, from the humanities. Our research group is interested in answering two tricky questions: What would an ethical AI be? And how can one guarantee that a given intelligent system will follow intercultural human ethical principles?

In this paper, we explore these two questions in two sections, in the hope of slightly broadening the perspective of the current

LAWS debate. In section 2, we show that there are discussions and research works currently being conducted worldwide on ethics and AI in general, which could shed a light on the particular debate on AI and weapons. Indeed, the task of ethics is to determine the elements that allow us to have and build intercultural dialogue. In section 3, we draw attention to the various forms of regulation beyond law or any kind of formal mechanism such as conventions. LAWS involve such complex issues, with critical consequences on humanity, that the debate and the solutions should not neglect all possible kinds of regulations.

2. Ethics for Artificial Intelligent Systems and how it affects LAWS

The more AI adoption advances in society, bringing socio-economic benefits, the more ethical questions are posed to governments, companies, and citizens on topics such as employment (certain human occupations will disappear, while new vacancies will be created); privacy (citizens leave digital tracks, but have little control over this data); and automation of decisions (which may be unfair and/or incomprehensible). On the latter, the most promising machine learning techniques, such as deep neural networks, involve complex models that cannot explain their decisions in a way that is understandable to the citizen. In addition, algorithms can incorporate bias against certain groups, as it is exemplified in the famous case of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, which tended to lengthen prison sentences for black people in the United States (Kirkpatrick, 2017; Spielkamp, 2017). Therefore, discussing the application of ethics in AI is becoming a hot topic in universities, enterprises, and governments.

2.1. Ethics and AI

From a practical dimension of the debate, ethics is not synonymous with morality. Ethics alludes to the collective, morality is about the behavior of an individual. Ethics, therefore, lends itself as a justification for the daily practices of people and organizations. If, on the one hand, ethics is, in philosophy, one of the three major fields of study, along with epistemology and metaphysics, on the other hand, it is a practice of uninterrupted reflection on choices, behaviors and decisions with the constant objective of the improvement of social life. Ethics is the collective debate in search of the corporate model that we, at present, want for our future, in this sense, it is a defense of intelligence, our dialogical and decision-making condition for the coexistence of the collective, the community, the groups, and that, to this day, persists in our socio-cultural practices, precisely in moments of greatest intellectual challenge.

Therefore, to discuss what an ethical AI would be, it is worth recognizing that ethics is a human concern and pursuit. Machines, even the ones presently considered intelligent, are far behind human "generalist intelligence". They do not comprehend the context of which they are a part. The IEEE (Institute of Electrical and Electronics Engineers) Ethically Aligned Design Manual (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019) warns about how misleading it may be to attribute to autonomous systems an anthropomorphic intelligence they do not possess.

Concerning this matter, Loh summarizes (Loh, 2019):

> It is currently assumed that technological developments are radically changing our understanding of the concept of and the possibilities of ascribing responsibility. The assumption of a transformation of responsibility is fed on the one hand by the fundamental upheavals in the nature of "the" human being, which are attributed to

> the development of autonomous, self-learning robots. On the other hand, one speaks of radical paradigm shifts and a corresponding transformation of our understanding of responsibility in the organizational forms of our social, political, and economic systems due to the challenges posed by robotization, automation, digitization, and industry 4.0. It is also expressed widely that, thanks to these circumstances, our modern mechanized mass society sets ultimate limits to responsibility, even opening up dangerous gaps in the possibilities of attributing responsibility.

The discussion on how to translate the principles of an intercultural human ethics to a machine, or to AI, is complex for many reasons: privacy concerns, responsibility for autonomous action, delegation of decision making, transparency, bias in collected and analyzed data, surveillance, and AI opacity. Isaac Asimov, in the 1950s, had already established firmly that robots, machines, and every other possible kind of artificial intelligence might be ***logical,* but not *reasonable***. And the inherent pondering that ethics brings about has to do with reasonability more than logic, as slight differences in context bring about completely different preferences and results. A good illustration as an example is the trolley problem and all of its posterior adaptations. (Ahlenius & Tannsjö, 2012; Goldhill, 2018; Judith Jarvis, 2008; Thomson, 1976; Waldmann & Dieterich, 2016; Ximenes, 2018)

2.2. Floridi's principles

Artificial Intelligence Ethics discussions have reached international spheres, and they are mapped in the AI Ethics Guidelines Global Inventory[5]. Another initiative worth mentioning

5 Available at <https://www.rri-tools.eu/-/ai-Ethics-guidelines-global-inventory>.

is the Algorithm Watch[6], an organization committed to evaluating and shedding light on the algorithmic decision-making processes that have compiled most of the AI Ethics manuals proposed so far.

In the current profusion of Ethics manuals and guidelines for AI, Prof. Floridi's AI4People framework emerges as the foundation for any serious discussion on the subject. (Floridi et al., 2018; Floridi & Cowls, 2019)

In this work, inspired by Bioethics principles, Floridi and colleagues from the Digital Ethics Lab at Oxford University propose five overarching principles highlighted as being the most important to be taken into account:

- **Beneficence** refers to a practice where the priority should maximize the benefit and minimize the loss. It may also be understood as promoting overall well-being, preserving dignity, and sustaining the planet. In some sense, institutions and states that have AI will be in a great position to create value if AI is used as a means to improve beneficence rather than diminish the well-being of citizens. "The prominence of beneficence firmly underlines the central importance of promoting the well-being of people and the planet with AI." (Floridi & Cowls, 2019, p. 4)
- **Non-maleficence** highlights precisely the main characteristic of the principle of beneficence. Thus, it establishes that the action must cause the least damage (action that does not do harm). In this sense we could propose, as examples, problems related to privacy, security, and misuse prevention for avoiding doing harm while trying to do good. As Floridi and Cowls comment, "it is not entirely clear whether it is the

6 Available at <https://algorithmwatch.org/en/>.

people developing AI, or the technology itself, which should be encouraged not to do harm." (Floridi & Cowls, 2019, p. 5)

- **Justice** establishes equity as a fundamental condition; thus, it is an ethical value in which each individual (agent) must be treated in accordance with what is morally correct and adequate and given what is due. The main characteristic of this principle is impartiality: acting with others disregarding their social, cultural, religious, financial and distinct aspects that may interfere negatively in the relationship. As put by Floridi and Cowls, "the diverse ways in which justice is characterized hints at a broader lack of clarity over AI as a human-made reservoir of 'smart agency'." (Floridi & Cowls, 2019, p. 6)
- **Autonomy** requires agents to have the skills and competencies to make decisions in a way that is respected for that. The vulnerability of agents, in specific or contingency circumstances, needs to be considered with respect to the decisions that will need to be made. In the sense of AI, Floridi and Cowls conclude that "the autonomy of humans should be promoted and that the autonomy of machines should be restricted and made intrinsically reversible, should human autonomy need to be protected or re-established."
- **Explicability (or Explainability)** is the need to understand and hold to account the decision-making processes of AI. This should be possible by providing intelligibility and responsibility to machine decisions through an accurate methodology in the core of the AI system that has implemented into itself a model of explicability. This is needed because there is a novel reality about AI: its functionalities and processes are invisible or unintelligible to almost all individuals. For Floridi and Cowls, this principle is possible, but also required, by

> "enabling the other principles through intelligibility and accountability." (Floridi & Cowls, 2019, p. 7)

Even though these principles seem abstract, requiring more precise guidelines for developers and decisors' daily activities, they do represent a good foundation for understanding what would constitute an ethical AI. This may be useful in the present context because LAWS are one of several specific applications of AI, and all such applications should ideally be adherent to the overarching principles of ethical AI. Beneficence, non-maleficence, and autonomy are more easily connected to the LAWS debate, and aspects related to each of these three tenets are mentioned throughout the 11 principles presented in the GGE/LAWS 2019 document (GGE LAWS, 2019). However, explicability is not explicitly mentioned in none such principles, and only (b), (d), and (h) are related to this vital aspect of building ethical AI through auditability, compliance, and accountability.

In traditional computer science, auditability has to do with the possibility of examining the source-code. However, machine learning, neural networks, and more modern and powerful AI techniques are black-box models by their very nature, making these systems harder to audit because they are not inherently explainable. The patterns found in data are often unclear to humans. It is also even more complex to determine accountability because part of the optimization and decisions is done according to parameters that AI engineers cannot directly control in detail. Therefore, extra effort has to be made by engineers and practitioners in order to provide clarity and explanations based on the model inputs and outputs. Research in the area of XAI (Explainable AI) is a growing field with more solutions and novel approaches being released every month. (Adadi & Berrada, 2018; Biran & Cotton, 2017; Gilpin et al., 2019; Powell et al., 2019) Typically, current solutions involve using statistical tools to probe for biases and building secondary computational or

mathematical models that approximate the system's behavior and optimization function.

There are different levels of explicability that can be provided, and they vary according to the criticality of the application and the level of expertise of the user. (Google, 2019) For instance, in a Netflix recommendation of movies to watch, a wrong recommendation does not carry heavy consequences, and the great majority of users are not specialists. Hence, the level of certainty of a recommendation is not as critical, and no explanation is given concerning how the AI system decided what to recommend. On the polar opposite, applications such as cancer diagnostic systems based on image detection carry heavy consequences for all people involved. Hence, they need further explanation to support the system output and diagnosis. LAWS are more similar to the second case, as they are employed in critical scenarios that have impact on life-or-death issues. Besides, explanation and auditing are normally carried out by experts in the area, who need more detailed information to make decisions or determine accountability. Users are not the only ones who benefit from explainability, as understanding the systems also carries benefits to legislators, legal departments, and engineers themselves, as it becomes possible to audit the models at some level, but this discussion is out of the scope for this paper.

Considering the benefits of having more explainable systems, we argue that a possible next step for the CCW/GGE Principles could be considering the need of providing some level of explainability and which the necessary metrics are to be used to determine whether a system should be deployed and used or not. The exact thresholds might be the subject of more debate, but it is important for practitioners creating these systems to understand the requirements, and that everyone involved understands what these autonomous systems take in as parameters to make decisions. Especially so because in some contexts (i.e. defending against attack), humans will be completely

out of the loop due to the need to respond promptly, and auditing and reviewing the decisions will be vital *post facto*.

2.3. Human role in machine-based decisions

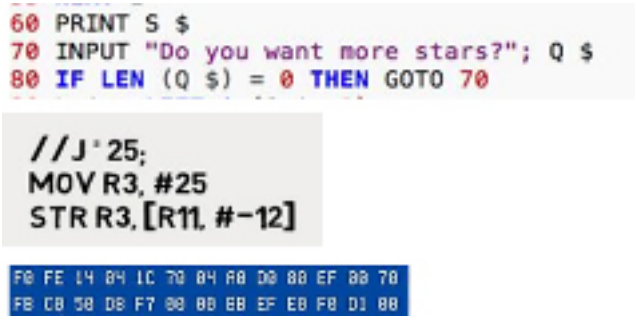
A central well-known issue in LAWS discussion is the role of humans in machine decision-making, commonly grouped into 3 categories, from high control to no one: Human in the Loop (HITL); Human on the Loop (HOTL); and Human out of the Loop (HOOTL). This discussion is, of course, not restricted to LAWS, even though, due to its criticality, it is especially applicable to this scenario. (Danks & Danks, 2013; Hoff & Bashir, 2015; Murphy & Woods, 2009)

In our research group, we have been trying to elicit some criteria currently used to decide the desired level of automation indecisions. We started by examining two domains, electricity distribution and Intensive Care Units, since these domains are highly regulated and involve risky decisions. We have found dozens of criteria. (Gilboy et al., 2011) For example, in US Emergency Rooms, these criteria are compiled in the ESI (Emergency Severity Index). Some of them may be useful or are already being used in the LAWS debate, such as: time to act (how much time is available for the decision); human factor (what the consequences of the decision on people's life are), environmental impact (what the consequences of the decision on the environment are); cost (what the overall cost of the automation is and what savings are generated by it); responsibility (how easy the identification of the responsibilities for the decision is); concurrency (how automation positions me in the face of competition), technical complexity (how complex and reliable the implementation of the automation is). The point here is to stress that it is important to establish a clear set of criteria on when to adopt fully automated decisions, how to do it, why do it, and who is able to do it. In the domains of electricity distribution and Intensive Care Units, this discussion seems to be more mature than in LAWS. We argue that, by

ensuring that every step taken in conceiving and building a system, from data collection to deployment, is ethical, we may equally ensure that a given intelligent system will have an overall ethical behavior as a consequence.

We can also explore the issue of computer autonomy from the engineering point of view. In computer science, we work with the notion of layers of abstraction. Each layer increases the level of abstraction, which means that, as we go to the upper ones, it is simpler to program a machine. The first layers are related to hardware, from the silicon substrate itself to the electronic boards. On top of that, there are the layers related to software, going from the machine code to Application Programming Interfaces, passing through assembly and programming languages. For those who are not familiar with these technicalities, imagine that, when one presses a brake or pushes the car's accelerator, this person does not need to know all the mechanical and electronic gears, mechanisms, and components involved in braking or accelerating the car. For the sake of clarity, this is an abstraction of the actual structure. Figure 1 illustrates three levels of software layers:

Figure 1: Example of three software layers. From the bottom, we have machine code, then assembly, then a simple programming language



What is AI from this point of view, after all? It is another layer of abstraction on top of programming. For instance, instead of programming the behavior of the machine, AI techniques allow the programmer to set only goals and rules, because the system has an embedded "inference engine" that knows how to start from a fact to deduce new facts according to the rules. Or the programmer can just give some examples of a given concept and let the machine learn the rules.

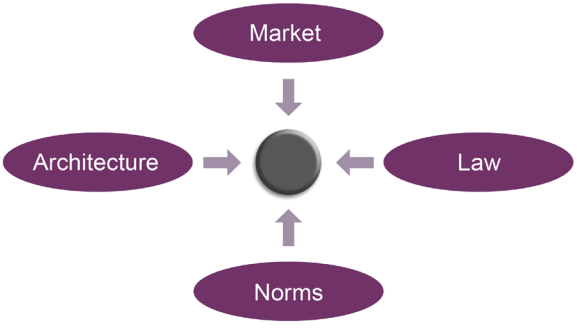
Thus, abstraction is necessary and the natural evolution of computer programming. The problem is that **the more abstract, the easier to program, but less control the programmer has over the machine**! So, the popular fear of losing control of machine decisions is not just a laic concern. Building ethical AI is a complex issue not only in philosophical terms, also fromthe technical point of view. (Russell, 2019)

3. Regulations for Artificial Intelligence and how it impacts LAWS

Half of the 11 principles proposed in the CCW/GGE 2019 document explicitly mention law, in particular compliance with International Humanitarian Law. There is no doubt that International Humanitarian Law is a fundamental reference to the debate on LAWS, and that it sets boundaries for what is allowed or prohibited. Moreover, the reference of this type of law in the context of a diplomatic debate is even more natural. However, if in the previous section we tried to broaden the perspective of the LAWS debate by pointing out that there are more AI artifacts than weapons in the ethical debate, in this section we want to emphasize that there are more forms of regulating AI artifacts than only laws. This is especially truer in the age we live, when technology is ubiquitous, and communication barriers have decreased, enabling a plethora of possible regulations.

Indeed, regulation is a broad concept. And in this context, we perceive regulation as any force or influence that changes the behavior of an agent, being able to limit or otherwise modify its actions. In order to establish our context and examples, we have adopted the Pathetic Dot framework proposed by Lawrence Lessig, which is very useful inexplaining and systematizing the discussion about regulation forces in the Internet era. (Lessig, 2006) Lessig explains that, from the point of view of someone or something that is being regulated, this entity is constrained by the inter-relations of four main forces, which are always balanced. Those forces are norms, laws, market, and architecture. The interaction between those forces can strengthen or undermine the influences of one upon the others, and their action is dynamic, changing across time. Figure 2 below illustrates the Pathetic Dot framework.

Figure 2: Pathetic Dot framework describing each regulating force



The specificities of each force are briefly explained below. We emphasize that the reader should keep in mind that many instruments and tools of regulation are an amalgam of different types and generate an influence in more than one sphere.

3.1. Laws

These are the most formal types of regulation, represented by constitutions, statutes, and legal codes. (Lessig, 2006) Laws are able to formally regulate and enforce not only the Pathetic Dot itself, but other forces as well. It can counteract norms or reinforce them with legal resources, limit market liberties, and define ideal architectures. While laws can be a highly effective form of regulation, in a democracy where representatives are elected, they also depend on several participants to write, redact, and push them forward from proposal to actual piece of legislation. Besides, laws are based on behavior that has already occurred, which by nature carries the consequence that law is implemented after it is necessary, as it is defined *post facto*. It does not have the intention or ability of foresight, and the phenomenon that it regulates must be well-described and understood. Code-based systems change too rapidly for lawmakers to describe and understand the phenomena they create in a timely manner. Therefore, even though the agents' compliance is supervised, and the law's punitive power exceeds that of any other means as well, it is important to realize the relevance of other regulatory forces, even if only as a means to compensate the inherent delay in the creation of applicable legislation. That is why, in this paper, we urge participants of the LAWS debate to open their minds to other possibilities of regulations, which could perhaps be as effective as formal laws.

3.2. Norms

Norms are essentially social constructs. They reflect relationships, culture, and behaviors of a given community. Norms are often informal and might never reach a written format, being instead based on the notion of what is acceptable or customary to do, then being an example of what should be done, in a cycle. Even so, some behaviors and habits can be recognized as especially desirable or in

need of standardization and may be registered in different ways. For instance, books on etiquette attempt to systematize norms. Best practice manuals for code maintainability and readability are sets of norms. ISO/IEC certifications are a way of auditing and asserting that certain norms are being followed, and they are valuable because society places value in such certifications. They differ from licenses, for instance, because they are not mandatory or enforced by the state; they are created by communities and maintained by private companies or the third sector. Norms are enforced by social pressure, and not complying with such terms might lead to loss of social capital and graver consequences, such as ostracism. In the examples presented above, none of the norms *must* be observed, but they might bear social consequences. For instance, not observing etiquette might mean not being invited for another dinner in the future; not complying with code maintainability and readability practices might mean losing the job; not having an ISO/IEC certification might mean losing clients to another company that has it.

Norms underlie all social relations and are not always explicit. An example of implicit norms would be Google's Project Maven for LAWS, which caused developers and AI engineers from Google to resign and walk out of their offices as a means to oppose the company's decision to participate in military projects. This happened because workers did not have the same expectations and moral code as the company, hence the fallout. This brought social and market repercussions to Google and spurred them to discontinue military collaborations. (Shane et al., 2018; Statt & Vincent, 2018)

Our research group elaborated, as a reference, a proposition of an Ethical AI Certification for companies in the private sector, based on the Great Place to Work and B Corporation certificates. Such certifications are recognized by society as a seal endorsing specific behaviors and qualities. This means they communicate value and are able to calibrate trust and expectations about a given product,

service, or company. We opted for a practical approach, and cross-referenced the five AI principles discussed by Floridi and introduced in the previous section with an extended CRISP-DM framework, one of the most popular Data Science frameworks, which describes a pipeline for creating data-based products covering elements from understanding business objectives to data preparation, to model training and deployment. (Wirth & Ripp, 2000) This produced a matrix, in which we are considering what a company should do to address, such as the issue of explicability in the data preparation phase. Answers are posteriorly audited against company evidence and depending on how the questions were answered. The company is awarded the Certification (which was dubbed CEIA – *Certificação em Ética para Inteligência Artificial*, in Portuguese; translated as Certification in Ethics for Artificial Intelligence). Some examples of the 48 questions from the reference questionnaire we are proposing are listed below:

- *Are the impacts of the positive effects of your system mapped in a clear and accessible way for all the company through the business targets and quarterly goals?*
- *Is it possible for humans to review and change decisions made by AI systems developed in the company that are used in critical settings?*
- *In data acquisition and preparation, is there a company-wide guideline for the target population to be represented equally, avoiding inherent data bias?*

These are just some examples, but they translate the ethical position of the company concerning its AI applications and the maturity of discussions and actions taken in relation to its positions. The CEIA then assumes a two-pronged approach: it guides processes internally, while communicating company priorities to employees and employers, and it also communicates to the outside world (e.g.: clients, citizens, third parties) what to expect from that company's

AI products and services, which can correspond to several economic advantages that are further discussed in the following subsection, "Market," the third regulating force we will explore.

3.3. Market

Markets are where economic exchanges take place. Simplistically, supply and demand curves meet at a marketplace and undergo adjustments to reach an equilibrium, which defines the existent quantity of a specific good or service, and the price at which it will be sold. Market equilibrium is dynamic, and these changes allow for market regulation of entities. Supply and demand curves may suffer shocks and be displaced, achieving new equilibria. Agents may also deliberately change their propension to buy or sell, also achieving new equilibria outside efficiency allocations in their original curves.

We can cite some examples of market regulation for AI. Buyers may boycott a company due to scandals, due to invasion of privacy or any kind of ethical issues. The lost reputation can be fatal for a company's survival. For instance, after Microsoft's facial recognition system was identified as being biased, they rapidly improved their training datasets and overall results. Even so, this piece of news harmed Microsoft's results in the quarter, and the company released a statement to investors explaining how biased or flawed systems can hurt the company's image, and why it is important to improve these models. (Gershgorn, 2019)

Another recent example, seen in the World Economic Forum 2020, is the decision of investment funds to condition their investments to projects that are committed to environmental, social and governance (ESG) issues. This positioning drives the market to a different direction. In the future, these premises may include Ethics for AI systems. Simultaneously, some government units stated that they will no longer purchase and deploy AI systems that cannot offer intelligible explanations for their decisions, in cases where

the decisions directly affect people's lives. For instance, New York City outlawed the use of black box models in the public sector, and Pennsylvania opted to have the state create its own recidivism risk assessment model, and have the code open for inspection, since it will not be proprietary to a private company. (Campolo et al., 2017; Pennsylvania Commission on Sentencing, 2019)

Insurance companies also play an important role in regulating markets. Suppose that the accuracy of weapons in distinguishing civilians from military targets is low. If a mistake is made, someone will have to pay a compensation for it, and the insurance company can be called upon to cover it. This will exert market pressure on the improvement of weapons accuracy, for instance. This will also exert pressure to create industrial benchmarks for LAWS, establishing quality standards for such systems.

3.4. Architecture

Architectural force has to do with the structure and the design of things, and how they can mold behaviors and regulate the way people operate. Unlike the other forces, architecture is an intrinsic aspect of the entity being regulated, a characteristic. An everyday example is airport benches and their armrests. These armrests are often static and cannot be elevated, which makes it more difficult, if not downright impossible, for a person to lie down and occupy multiple seats at once. This has to do with the bench's *architecture*, its structure. The act of lying down could be regulated in multiple ways, such as (a) by outlawing the act and arresting the person (which may sound preposterous in an airport setting, but it happens in park benches all over the world, where the homeless might be arrested for loitering); (b) by normatively embarrassing the person through insistent glares and disapproving looks or, (c) more lightly, by putting signs requesting that people think of other tired passengers and do not occupy more than one seat at once; or yet, (d) by applying

a monetary fine if the person is caught lying down, combining law and market forces to regulate chair occupancy in airports. All these have the same goal, but we consider that the architectural approach is more direct and more likely to work, not only for airport benches but also for AI systems. Indeed, some architectures can be easily changed (e.g. the items displayed on a software, such as the first screen of a mobile phone), and others are more permanent (e.g. the https protocol to safely transfer data packages online).

Changing the architecture implies changing what something is as well as how it should operate. It is deeply connected to the concept of feasibility and what a system consists of (i.e. code is the building brick of software). However, changes in architecture can also be applied to less concrete things, such as processes. Changing the steps of a pipeline generates structural changes and new demands not only throughout the process, but also in the final result. For instance, the inclusion of automated testing and quality assurance steps in software development and industrial pipelines spurred practical changes in tasks and processes, and had direct results on final systems and goods. Therefore, there are multiple ways to influence how things are through architectural changes.

Similarly to our certification proposal, our research group has also conceived a Consumer Artificial Intelligence Information Leaflet (CAII), similar to Patient Information Leaflets (PILs) that accompany medicines. Drugs have different compositions and purposes, but they have uniform processes, tests, and standards the pharmaceutical company must follow to get them approved. (US Food and Drug Administration, 2019) We find that this heavily resonates with AI issues, as we were able to draw a parallel with the drug approval process (based on material from the FDA-USA, TGA-Australia, and ANVISA-Brazil). In all cases, it is necessary to undergo four phases: application, clinical studies, approval process, and post-market tests. In the application phase, the company must

provide the basic data about purpose, application, dosage, and overall population characterization, which is highly applicable to AI products. The clinical studies encompass the effects of the drug on the human body, its efficacy, safety, and side effects. These can be adapted to the AI context and consider safety, security, biases, and a study of social impacts. In the approval process, for both cases, results are audited and checked for compliance with current applicable laws. Finally, in the post-market tests, the efficacy and consequences of the product are tested on a large scale. In the end, all the highlights are condensed and provided in a single leaflet that is freely circulated and to which everyone can have access. Even though this approach blurs the lines of individual regulation forces, covering laws, norms, and market, it also helps understand *how* to build an ethical AI system and guides architectural decisions and processes.

Architecture is the main factor we influence in computer science, and this is what we consider a key for building ethical machine learning systems, LAWS included. We believe the most critical changes and decisions for ethical systems must be made, by design (as stated in Principles (c) and (g) of the CCW/GGE 2019), before any final product exists. The most efficient regulation from the standpoint of a system is one that imposes constraints while the system is being created, as it limits from the very beginning what a system can or should do. Applicable product constraints will be elicited according to principles discussed in intercultural forums, such as the CCW/GGE forums.

Considering LAWS, we might take into account, during the construction of the system, specifications that comply with ethical standards. For instance, if one is creating land mines that should not be activated by a person, only war tanks, the weighing sensors used must be able to identify and differentiate weights, so the mine is only triggered in the correct context. These sensors must be embedded in the device during the process of its construction,

before the product is ready. Similarly, one might create a certainty threshold for image recognition, so that if an attack is conducted with LAWS and the target is unclear, the weapon remains locked and cannot take further action. For instance, it may be acceptable to proceed with 80% certainty of what the target is—military personnel or civilian, vehicles, buildings, among others, and this information is automatically calculated by any machine learning model. For systems with a human in the loop, this level of certainty (i.e. precision/recall) can be shown on screen so decisions are made with the correct information; for systems with humans out of the loop, this information can be automatically taken into consideration as a condition to initiate an attack.

The crucial question is *how* to do that, as the relevant aspects must be considered beforehand to create a system known as "ethical by design." This recognizes that observing ethical principles cannot be incidental, but instead must be planned and built into the system itself. Floridi himself explains that this ethical design is about an approach model that can protect and promote the aforementioned ethical tenets (specifically the AI decision-making processes), thus incorporating them from the beginning into the design specifications, functional and non-functional requirements of technologies (e.g.: AI, robots, etc.), procedures, practices or infrastructures. Our aim here is not to document every single ethical consideration for an AI project, but to consider and propose, as a debate that should be self-evident, that an AI project ought not to advance the proliferation of unchecked and unaccountable weapons. (Taddeo & Floridi, 2018)

4. Final remarks

In this article, we have discussed the worldwide debate that has been ongoing for a few years concerning what would be considered an ethical AI and how to achieve it. The core discussion is not about whether LAWS should be allowed or banned, but instead about

how they are part of a broader scenario of AI ethics and systems, and where to look in order to advance the discussion in practical ways. We have presented the five guiding principles for ethical AI, and argued that the pillar of explainability should be directly considered in the CCW/GGE Principles, since it allows for a better understanding of the AI systems, what they can do, how likely they are to achieve specific goals, as well as establishing the ground zero for any discussions on compliance, auditability, and accountability that are vital for LAWS. We also addressed the discussion on the role of human beings in machine decision automation, remembering that adopting AI techniques and tools simplifies programming, but also implies a certain loss of control.

We have also broadened the discussion on regulation under the lenses of the Pathetic Dot framework for the Internet era and code-based products. Laws are one of the four determining forces that regulate any entity, but it is possible to incentivize other behaviors and the production of accountable systems through normative, market, and architectural forces. We posit that architecture is a strong and often overlooked regulatory force as it depends on deep technical knowledge, but it also corresponds to reliable results in a myriad of scenarios beyond LAWS and beyond AI applications, as it shapes the very structure and capability of a system beforehand. We have illustrated our argument with some practical tools for regulating AI systems we have created in our research group on Ethics and AI; these were the Certification for Ethical AI and the Consumer AI Information Leaflet. Such tools could be applied to the LAWS scenario as well.

The debate on ideal LAWS and ethical artificial intelligence have much in common and would benefit from sharing more common ground. Furthermore, alongside the intercultural forums, we also need interdisciplinary forums where we can unite legislators, thinkers, practitioners, and idealists to define what to pursue for

the future of humanity alongside artificial intelligence. Only then will we be able to identify a wide range of approaches, opting for the more efficient ones that comply with our ethical principles and moral values.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*(c), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahlenius, H., & Tannsjö, T. (2012). Chinese and Westerners Respond Differently to the Trolley Dilemmas. *Journal of Cognition and Culture*, *12*(3–4), 195–201. https://philpapers.org/rec/AHLCAW

Biran, O., & Cotton, C. (2017). Explanation and Justification in Machine Learning: A Survey. *IJCAI Workshop on Explainable AI (XAI)*, *August*, 8–14.

Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). *AI Now 2017 Report*.

Danks, D., & Danks, J. H. (2013). THE MORAL PERMISSIBILITY OF AUTOMATED RESPONSES DURING CYBERWARFARE. *Journal of Military Ethics*, *12*(1), 18–33. https://doi.org/10.1080/15027570.2013.782637

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and

Recommendations. *Minds and Machines*, *28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gershgorn, D. (2019). *Microsoft warned investors that biased or flawed AI could hurt the company's image*. Quartz. https://qz.com/1542377/microsoft-warned-investors-that-biased-or-flawed-ai-could-hurt-the-companys-image/

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, 80-89. https://doi.org/10.1109/DSAA.2018.00018

Goldhill, O. (2018, February). Philosophers are building ethical algorithms to help control self-driving cars. *Quartz*. https://qz.com/1204395/self-driving-cars-trolley-problem-philosophers-are-building-ethical-algorithms-to-solve-the-problem/

Google. (2019). *People + AI Guidebook*. https://pair.withgoogle.com/

Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE LAWS). (n.d.). *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*. https://undocs.org/en/CCW/GGE.1/2019/3.%0A

Hoff, K. A., & Bashir, M. (2015). Trust in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Judith Jarvis, T. (2008). Turning the Trolley. *Philosophy & Public Affairs*, *36*(4), 359–374. https://doi.org/10.1111/j.1088-4963.2008.00144.x

Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Communications of the ACM*, *60*(2), 21–23. https://doi.org/10.1145/3022181

Lessig, L. (2006). *Code: And Other Laws of Cyberspace, Version 2.0* (2nd ed.). Basic Books. http://codev2.cc/download+remix/Lessig-Codev2.pdf

Loh. (2019). Responsibility and Robot Ethics: A Critical Overview. *Philosophies*, *4*(4), 58. https://doi.org/10.3390/philosophies4040058

Murphy, R. R., & Woods, D. D. (2009). Beyond Asimov: The Three Laws of Responsible Robotics. *IEEE Intelligent Systems*, *July/August*, 14–20.

Pennsylvania Commission on Sentencing. (2019). *Adopted Sentence Risk Assessment Instrument*. http://pcs.la.psu.edu/guidelines/adopted-sentence-risk-assessment-instrument

Powell, A., Joshi, A., Carfantan, P.-M., Bourke, G., Hutchinson, I., & Eichholzer, A. (2019). *Understanding and Explaining Automated Decisions*.

Russell, S. J. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*.

Shane, S., Metz, C., & Wakabayashi, D. (2018). How a Pentagon contract became an identity crisis for Google. *The New York Times*. https://www.nytimes.com/2018/05/30/technology/google-project-maven-pentagon.amp.html

Spielkamp, M. (2017). Inspecting algorithms for bias. *MIT Technology Review*. https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/

Statt, N., & Vincent, J. (2018). Google pledges not to develop AI weapons, but says it will still work with the military. *The Verge*.

https://www.theverge.com/2018/6/7/17439310/google-ai-ethics-principles-warfare-weapons-military-project-maven

Taddeo, M., & Floridi, L. (2018). Regulate artificial intelligence to avert cyber arms race. *Nature*, *556*(7701), 296–298. https://doi.org/10.1038/d41586-018-04602-6

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically Aligned Design A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems First Edition The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*.

Thomson, J. J. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist*, *59*(2), 204–217. https://doi.org/10.5840/monist197659224

US Food and Drug Administration. (2019). *Artificial Intelligence and Machine Learning in Software as a Medical Device*. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Waldmann, M. R., & Dieterich, J. H. (2016). Throwing a Bomb on a Person Versus Throwing a Person on a Bomb: Intervention Myopia in Moral Intuitions. *https://Doi.Org/10.1111/j.1467-9280.2007.01884.X*, *18*(3), 247–253. https://doi.org/10.1111/j.1467-9280.2007.01884.x

Wirth, R., & Ripp, J. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, *24959*, 29–39. https://doi.org/10.1.1.198.5133

Ximenes, B. H. (2018). Non-intervention policy for autonomous cars in a trolley dilemma scenario. *AI Matters*, *4*(2), 33–36. https://doi.org/10.1145/3236644.3236654

Panel 1 - Human-machine interaction and human control

Geber RAMALHO

Background

- Electronic engineer (1988)
- PhD in Artificial Intelligence – Paris VI (1997)

Currently positions

- Professor in Computer Science Center - UFPE
- Chairman of the board of CESAR Institute

Interests

- AI for art and entertainment
- Ethics and AI
- Innovation and entrepreneurship

What would be an ethical AI?

How to guarantee that a given intelligent system will have an ethical behavior?

Luciano Floridi's Principles

1. **Beneficence**: promoting well-being, preserving dignity and sustaining the planet
2. **Non-maleficence**: privacy, risk and misuse prevention, "capability caution"
3. **Autonomy**: the power (of the user) to decide (or not)
4. **Justice:** promoting prosperity and preserving solidarity
5. **Explicability** (giving machine decisions intelligibility and responsibility

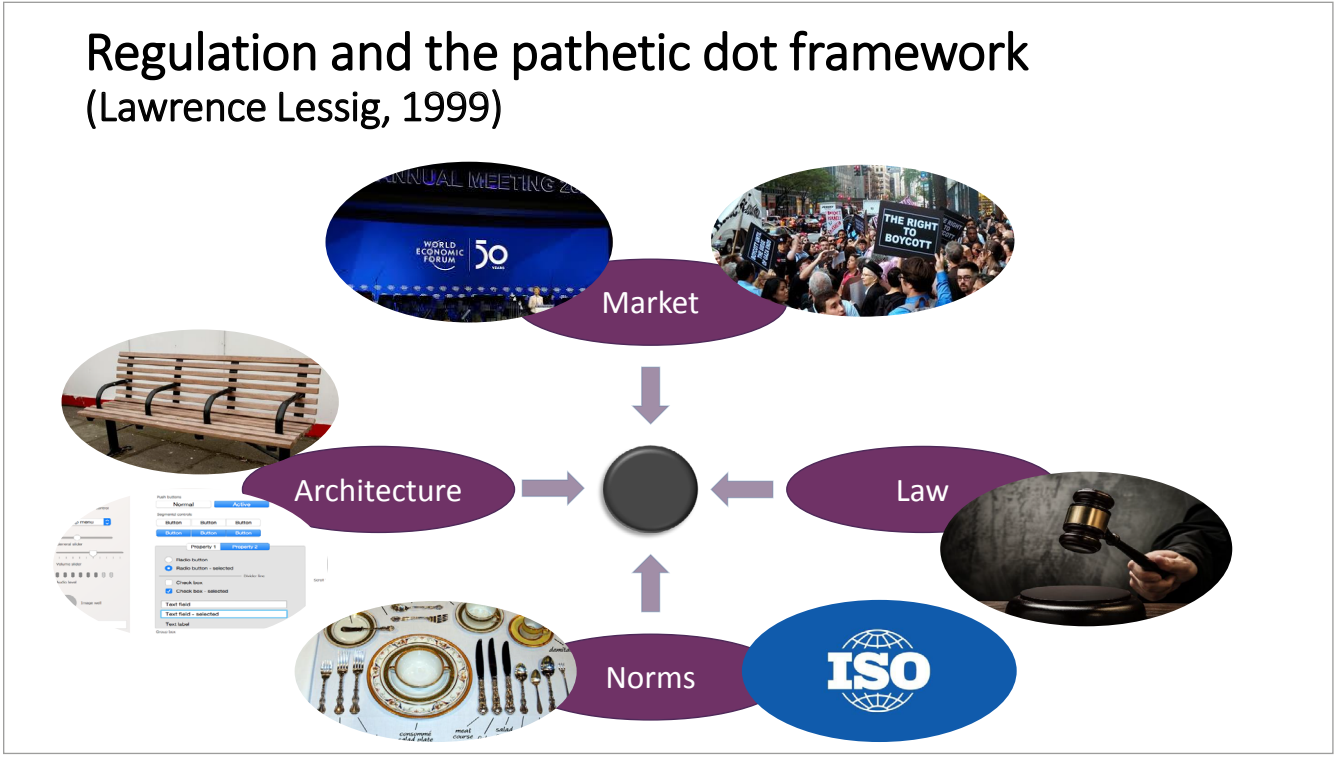
Ban LAWS!

- **In some cases?**
- **Under certain circumstances?**
- **For some weapons?**

How to “limit the damage”?

Which are the adoption critera, processes, responsibilities?

Some insights from our research group on Ethics and AI at UFPE


Regulation and the pathetic dot framework
(Lawrence Lessig, 1999)
WORLD ECONOMIC FORUM
50
THE RIGHT TO BOYCOTT
Market
Architecture
Law
Norms
ISO

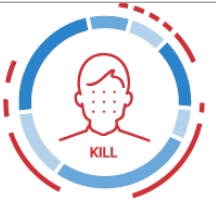
Law: Criteria for the adoption of fully automated AI

- Preliminar work
- Identify criteria for adopting HOOTL (Human out of the loop) approach in 3 (regulated or requiring regulation) domains
  - Intensive Care Unities
  - Electricity distribution
  - Lethal Automated Weapons
- Compare them looking for convergence

Intensive Care Unities

Case	Description	Examples	Interaction
Resuscitation	Immediate intervention to save life	- Cardiac arrest - Massive bleeding	HITL
Emergency	High risk of deterioration (leading to death) or signs of critical problems	- Chest pain (cardiac) - Asthma Attack	HITL
Urgent	Stable but requires multiple resources for diagnosis and treatment (laboratory tests, X-rays, tomography, etc.).	- Abdominal pain - High fever with cough	HOTL
Slightly urgent	Stable requiring few resources (a simple X-ray or sutures).	- Simple laceration - Pain when urinating	HOTL
Not urgent	Stable without need for resources beyond the prescription	- Abrasion - Renew medicine	HOOTL

LAWS



- Target precision (distance) α HOOTL
- Reponsibility/Explicability α HOOTL
- Damage Extent 1/α HOOTL
- Context/Environment complexity 1/α HOOTL
- Dignity (human as target) 1/α HOOTL

Comparison: criteria influence for adopting HOOTL

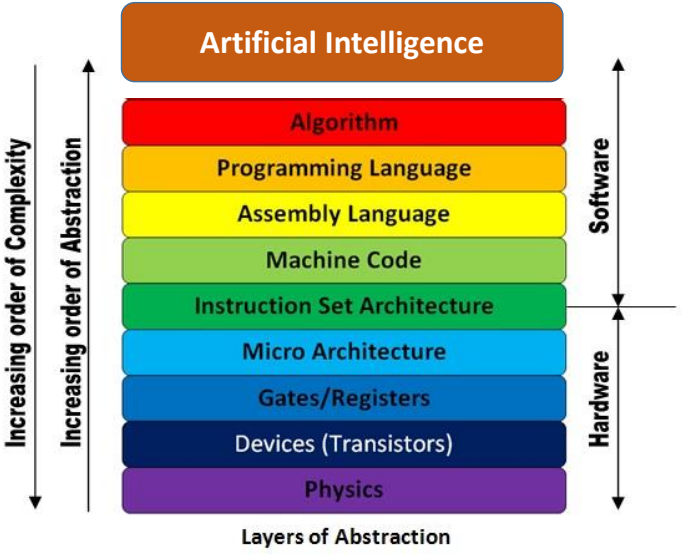
Time to act	α	1/α	1/α
Impact on people	1/α	α	1/α
Cost		α (operation)	α (troop life)
Responsibility	1/α	1/α	1/α

Market: certifications

- Ethical AI for enterprises (similar to the B-system and “great place to work”)
  - CRISP-DM process vs. Floridi’s principles => 48-questions questionaire

	Business understanding	Data understanding	Data acquisition	Data preparation	Modeling	Evaluation	Deployment	Observation in the wild
Beneficence: Promoting Well-Being, Preserving Dignity, and Sustaining the Planet	Ethical business goals + impacts	Guarantee that all populations are represented equally in the data	unpleasant data request + compliance	unpleasant data exclusion, stratified samples, data balancing	Transfer learning; Escolher modelos energeticamente eficientes, data parameterization	Assess whether the model has deviated from initial beneficent goals during the development process		
Non-maleficence: Privacy, Security and "Capability Caution"	Risk planning; System legal compliance	using data with potential misuse	Using data only when under affirmative consent; Deleting data and erasing traces when consent is revoked; Not buying personal user data	Data protection through ISO 27000 serie	Design rigorous testing processes for applications that deal with sensitive data	Impacts evaluation, regression analysis		
Autonomy: The Power to Decide (Whether to Decide)	Automation impact assessment; Critical areas / decisions;	social discrimant propagation	Automatic dataset increment	Probing correlations in data, removing sensitive data and their proxies	unbalanced errors considerations, local minimium use, isolated examples exclusion	Need for model efficiency monitoring		
Justice: Promoting Prosperity and Preserving Solidarity	Doesn't allow for poverty + social entrapment;	Legal compliance	unbalanced data acquision	Pertinent demographic groups are represented in equal proportions on the training and testing datasets	Design tests aimed at detecting unfair treatment of demographic groups, analyze the robustness	Assess whether the model discriminates against demographic groups	Não contrariar restrições locais.	
Explicability: Enabling the Other Principles Through Intelligibility and Accountability	Business model visibility; Business model updated.	Data correlation and importance	Data lake acquisions	Data preprocessing record	Surrogate models;	Document mining process in final report; confusion matrix analysis	Ethical requirements (RNF) must be materialized and implemented.	Ethical Commite

Architecture: abstract layers in computing



Rules + reasoning > goals > learning

```
60 PRINT S $
70 INPUT "Do you want more stars?"; Q $
80 IF LEN (Q $) = 0 THEN GOTO 70
```

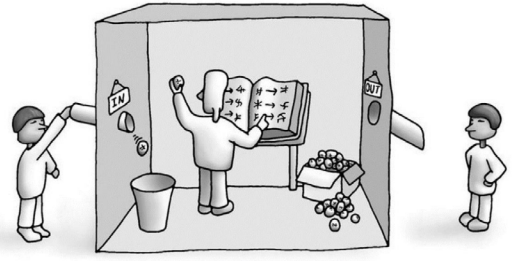
```
//J˚25;
MOV R3, #25
STR R3, [R11, #-12]
```

```
F0 FE 14 04 1C 70 04 A0 D0 00 EF 00 70
FB C0 50 D8 F7 00 00 BB EF E0 F0 D1 00
```

The more abstract, the easier to program, but less control you have!

AI limitations: reasoning

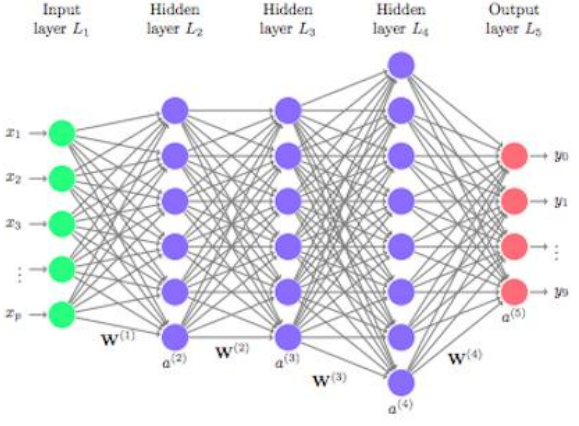
A) $\forall$ x,y,z Americano(x) $\wedge$ Arma(y) $\wedge$ Nação(z) $\wedge$ Hostil(z) $\wedge$ Vende(x,z,y)
$\Rightarrow$ Criminoso(x)
B) $\forall$ x Guerra(x,USA) $\Rightarrow$ Hostil(x)
C) $\forall$ x InimigoPolítico(x,USA) $\Rightarrow$ Hostil(x)
D) $\forall$ x Míssil(x) $\Rightarrow$ Arma(x)
E) $\forall$ x Bomba(x) $\Rightarrow$ Arma(x)
F) Nação(Cuba)
G) Nação(USA)
H) InimigoPolítico(Cuba,USA)
I) InimigoPolítico(Irã,USA)
J) Americano(West)
K) $\exists$ x Possui(Cuba,x) $\wedge$ Míssil(x)
L) $\forall$ x Possui(Cuba,x) $\wedge$ Míssil(x) $\Rightarrow$ Vende(West, Cuba,x)



M) Possui(Cuba,M1) - *Elimination of existential quantifier and the conjuction in K*
N) Míssil(M1) - instantiation
O) Arma(M1) - *Modus Ponens from D e N*
P) Hostil(Cuba) - *Modus Ponens from C e H*
Q) Vende(West,Cuba,M1) - *Modus Ponens from L, M e N*
R) Criminoso(West) - *Modus Ponens from A, J, O, F, P e Q*

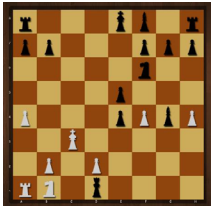
AI limitations: explicability

- Sometimes decisions cannot be explained!



AI limitations: one task-oriented

- AI has good performance in narrow application domains



The story of Carlsen winning the "double," getting the triple crown and finishing the year as the world champion and world number-one in standard, rapid and blitz is big. However, the incident on the last day in his game with **Alireza Firouzja**, who lost on time and whose protest was rejected, boosted the comments even further, and this story just makes it into the top-10! **208 comments** (at the time of writing!).

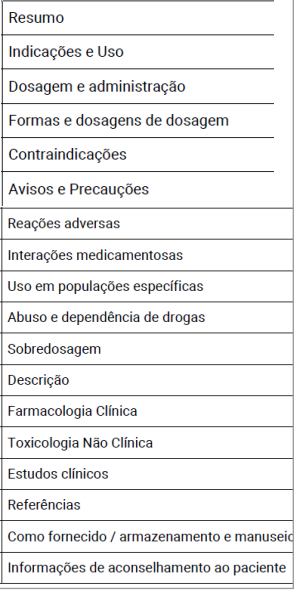
Architecture: AI limitations must be tracked and stated clearly

GGE principle (g) "Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems"

How to translate this into a practical measure?

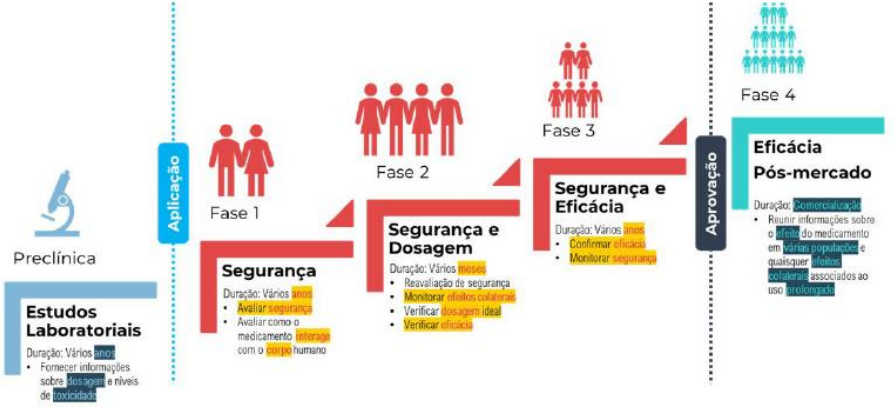
Consumer Artificial Intelligence Information (CAII)

- A parallel with pharmaceutical industry!
  - FDA (US), ANVISA (Brazil), TGA (Australia)
- Concerning drugs
  - It is approved through a long process of tests
  - We know a lot of things (18 items): eficacy, side-effects, constraints, dosage...
  - The Prescribing Information is a contract



Consumer Artificial Intelligence Information (CAII)

- The research, approval and deployment process for AI systems



Technology is part of the problem, but may be part of the solution!

Algorithms for mitigation of bias

- Optimized Preprocessing (Calmon et al., 2017)
- Disparate Impact Remover (Feldman et al., 2015)
- Equalized Odds Postprocessing (Hardt et al., 2016)
- Reweighing (Kamiran and Calders, 2012)
- Reject Option Classification (Kamiran et al., 2012)
- Prejudice Remover Regularizer (Kamishima et al., 2012)
- Calibrated Equalized Odds Postprocessing (Pleiss et al., 2017)
- Learning Fair Representations (Zemel et al., 2013)
- Adversarial Debiasing (Zhang et al., 2018)
- Meta-Algorithm for Fair Classification (Celis et al.. 2018)

Consumer Artificial Intelligence Information (CAII)

- 41 information items covered
- About AI
  - Intended use
  - Explanations
  - Model resource
  - Algorithms
  - Training data
  - Training environment
  - Optimizartion goals
  - User intarface
- About data
  - Sensors and sources
  - Actuators and outputs
- Legal Information
  - Lead programmer
  - Registration
  - Developed by
  - Consumer contact
  - Impact report
  - ...

Consumer Artificial Intelligence Information (CAII)

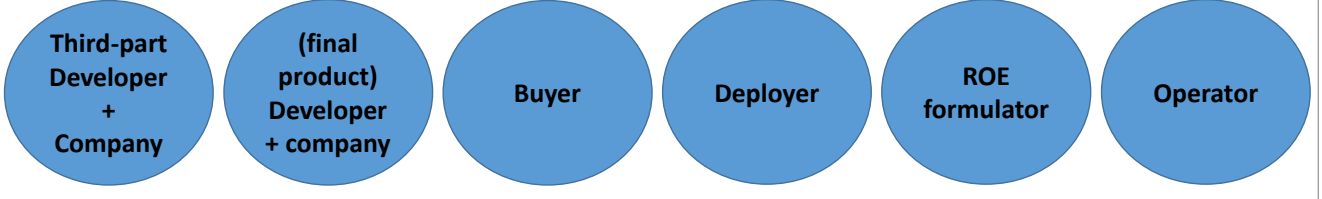
- Consumer information
  - What should I know to use the system?
  - How my data will be used?
  - Where and for how long my data will be stored?
  - Who my data will be shared with?
  - When my data will be shared?
  - ...

Main messages

- Fully automated weapons may perahps be inevitable, but risks should be cotrolled and technology + regulation can help
- Law is not the only possible regulation, and sometimes not the best one
- It is worth looking at what is being discussed in ethics and AI in general

Final remarks on GGE's principles

- (b) **Human responsibility** for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines.
- **Who? Define clearly the stakeholders!**



Panel 2: International Law, Including IHL, on LAWS: Is There a Need for a New Protocol?

Moderator: Pamela Moranga Quezada

Delegation to the United Nations in Geneva (Chile)

Thank you very much.

Thank you to our organizers today.

Welcome back from lunch. I hope you enjoyed it, and I hope you don't fall asleep.

We are back to start panel number 2: International Law, Including IHL, on LAWS: Is There a Need for a New Protocol?

So this last question is our teaser.

As we know, I am not going to bore you with details because you have heard long and lengthy presentations in the morning.

Lethal Autonomous Weapons Systems are not specifically regulated by any international humanitarian law treaty.

However, if there is one thing that discussions through the years, informal and formal, in Geneva, have been saying and stating as a fact is that they do not operate in a legal vacuum, they must be used in accordance with International Humanitarian Law or in other relevant legal frameworks.

That said, it has also been acknowledged that LAWS are something different; for starters, they are not a distinct singular weapon, it is rather a function of autonomy in a weapon. This unique feature sets it apart, and goes beyond the traditional dichotomy, as we know it, of conventional and unconventional weapons, let alone the ethical dimension to it, even if they comply with IHL, would we still delegate taking lives from a human being to a machine, which is also an important aspect of this.

To help us navigate through these waters, we have here an expert panel, and I will introduce you to them.

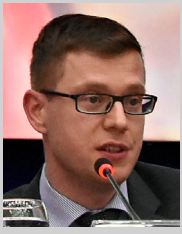
We have Konstantin Vorontsov, Head of Division of the Department for Non-Proliferation and Arms Control, Ministry of Foreign Affairs of the Russian Federation;

Kathleen Lawand—who will be joining us via Skype, so please be patient with us—Head of Arms Unit and Legal Division of ICRC (International Committee of the Red Cross);

Ambassador Michael Biontino, Senior Special Adviser from the German Federal Foreign Office;

Bonnie Docherty, lecturer on law at Harvard Law School;

And last but not least, Ambassador Thomas Hajnoczi, Director for Disarmament, Arms Control and Non-proliferation from the Ministry for Europe, Integration and Foreign Affairs of Austria.



Talking Points

Konstantin Vorontsov
Department for Non-proliferation and Arms Control (Russia)

Colleagues,

As you all know, the Russian Federation participated actively in all meetings of the GGE-LAWS in the framework of the CCW from its outset. We highly appreciate the GGE's work and its fruitful outcomes. In particular, we welcome the adoption and endorsement by consensus of two substantive reports in 2018 and 2019 containing the 11 Guiding Principles in the context of LAWS. These deliverables confirm that the CCW-GGE on LAWS is the optimal venue for considering the LAWS issues.

One of the key agreed guiding principles is that the existing provisions of IHL are fully applicable to LAWS. In addition, there are no indicators that the legal norms need to be adapted to the specificity of these weapons systems. We have all the necessary instruments

to regulate LAWS, as it is a matter of responsible implementation of the IHL norms and principles.

The IHL provides the necessary basis for the development and employment of such systems. I would like to remind you that the Additional Protocol I (AP-I) to the 1949 Geneva Conventions, which introduced the principles of proportionality by prohibiting indiscriminate attacks, obliges states to take precautionary measures for the sake of protecting civilian population. Its Article 36 also obliges member states in the study, development, acquisition or adoption of a new weapon, means or method of warfare, to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law.

Many states including Russia, fully comply with its relevant obligations. For this reason, it is unnecessary to elaborate a mandatory mechanism for "legal reviews," designed especially for LAWS. It is more important to universalize AP-I and urge states to withdraw their reservations made after ratifying this IHL instrument.

Therefore, we cannot agree with the view related to the elaboration, within the GGE framework, of any legally binding instrument on LAWS or moratorium on the development and use of technologies designed to create these systems. Due to many factors, we also consider it to be premature to discuss a "code of conduct" in relation to LAWS.

We proceed from the widely shared understanding that the discussion on LAWS should not focus only on the issue of various legal options, but must continue in a comprehensive and balanced manner and in full accordance with the subject and objectives of the CCW, the GGE terms of reference, and the agenda. Logically, that prioritization of one issue above others will provoke increasing practical difficulties.

Thus, it seems necessary to stimulate the dialogue on the issues of LAWS characterization, its military application, and human control. Having a common definition of LAWS is to be one of the main priorities for the GGE, otherwise each state would have its own interpretation of LAWS as well as understandings and guiding principles. Such a situation would lead to unpredictable consequences, concerning their practical implementation. There could be misunderstandings, causing the division of weapons into "good" and "bad" ones. It is better to avoid such a scenario. In addition, the definition of LAWS is subject to further work related to key aspects of this type of weaponry, such as the concept of autonomy, critical functions, human control, predictability, reliability, and so on.

Furthermore, the lack of ready-to-work samples of such weapons systems remains an issue in the discussions on LAWS. Notwithstanding the precedents for reaching international agreements that establish preventive measures, including a ban on prospective types of weapons, that approach can hardly be considered an argument—just like "one size fits all" for taking preventive, prohibitive or restrictive measures against LAWS. LAWS are a far more complex and wider class of weapons of which the current understanding of humankind is rather approximate. Before taking any practical steps, we all need to understand all the aspects of such weapons systems and the positive or negative consequences of their possible application. It is our common objective to avoid any harm to scientific and technical progress in the spheres of information technologies, artificial intelligence, peaceful robotics, etc.

We agree that human control is fundamental to provide a predictable application of the machine. Thus, as advanced as it may be, any autonomous system cannot perform its functions without a human behind it. Hence, the human who operates or programs the robot's system and orders the use of LAWS should take the responsibility for the use of LAWS.

Russian experts remain convinced that the CCW is an optimal forum to consider the matters related to LAWS. Indeed, this Seminar is a unique instrument and provides a format where deliberations are held and issues are discussed on the basis of a reasonable balance between humanitarian concerns and legitimate defense of the interests of states. This particular feature provides a practical opportunity to analyze such a contradictory and controversial subject as LAWS with realism and due prudence. Thereby, we reaffirm our determination to continue the active engagement in the work of the GGE under the subject and objectives of the CCW to obtain further results, and call on other partners to do the same.

International Law, Including IHL, on LAWS: Is There a Need for a New Protocol?

Kathleen Lawand
Head at the Arms Unit – Legal Division
International Committee of the Red Cross

Introduction

Let me begin by thanking Brazil, in particular Ambassador Candeas and his team, for organizing this very important seminar. And I thank the moderator of this panel, Ms Pamela Moranga of Chile, for her kind introduction.

For the ICRC, the human role in the use of force is indeed the fundamental question at the heart of the humanitarian, legal, ethical and societal issues raised by autonomy in weapon systems.

We have called on States in the CCW's GGE-LAWS to work towards common understandings on the elements of human control over the critical functions of weapon systems needed for legal compliance and ethical acceptability.

To help guide work in this respect, in June 2020 the ICRC and SIPRI published a report on *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*. We are grateful to the Netherlands, Sweden, and Switzerland for funding this project. To ensure compliance with IHL (also known as the law of armed conflict) and ethical acceptability, the report recommends three types of control measures: on the weapon's parameters, on the environment of their use, and in relation to human-machine interaction. I will provide a bit more detail about these control measures later.

But first I will focus on the application of IHL to autonomy in weapon systems.[1] I will make three main points:

- Existing IHL already provides some limits on the use of autonomous weapon systems (AWS).
- However, the unique characteristics of AWS—and in particular the unpredictability of their effects—presents unique challenges and difficulties in interpreting and applying the relevant IHL rules, which do not find clear answers in existing IHL. Ultimately, these challenges raise the question of whether there is a need to clarify IHL or develop new rules (I should also mention here that the limits dictated by ethical considerations may go beyond those found in existing IHL rules, notably with respect to anti-personnel systems).
- Some practical constraints on AWS might be derived from existing IHL, to ensure that human control is maintained at a meaningful level.

1 The IHL analysis in this paper is based notably on the section on autonomous weapons of the ICRC's report to the 33rd International Conference of the Red Cross and Red Crescent in December 2019, "IHL and the Challenges of Contemporary Armed Conflicts."

Characteristics of autonomous weapons that raise concerns under IHL

The ICRC understands autonomous weapon systems (AWS) as weapons that select and apply force to targets without human intervention. The weapon self-initiates (or triggers) a strike in response to what it senses in the environment, based on a generalized "target profile."

In the use of classical (non-autonomous) weapon systems, these functions are carried out by humans: the user chooses the specific target(s) and knows the location and timing of strike(s) when launching an attack. The central consequence of autonomy in the critical functions of weapon systems is a change in—indeed a diminishing or erosion of, the role played by humans in the use of force. A commander activating an AWS may know what type or class of target the AWS is intended to strike, but, to varying degrees, depending on the circumstances, the commander knows neither the specific target(s) nor the exact timing and location of strikes that will result.

This uncertainty means that the consequences of an attack using an AWS will always be unpredictable to a degree, especially in dynamic environments, putting civilians and other protected persons, as well as civilian objects, at risk and raising significant challenges for IHL compliance.

What are the limits that IHL already imposes on autonomy in weapon systems?

The short answer to this question is that IHL rules on the conduct of hostilities—distinction, proportionality, precautions in attack—must be complied with by those persons who plan, decide on and carry out attack. The assessments required by IHL rules involve evaluative and contextual judgements, for which humans are

responsible and accountable. These context-based human judgments cannot be substituted with machine, sensor, or software functions. Existing IHL rules requires commanders/operators of AWS to retain a level of human control over weapon systems sufficient to allow them to make the required context-specific judgments.

Let us look at this a bit more closely.

Like any weapon, AWS must be capable of being used and must be used in accordance with IHL rules on the conduct of hostilities—notably the rules flowing from the principles of distinction, proportionality and precautions in attack. These include:

- The requirement to distinguish at all times civilians and civilian objects from combatants and military objectives, and to direct attacks only at the latter. It should be noted in this respect that it is allowed to direct attacks against civilians that are directly participating in the hostilities, for the duration of such direct participation.
- The prohibition to attack combatants that are *hors de combat* because captured, wounded or surrendered.
- The prohibition to launch indiscriminate attacks—i.e. those of a nature to strike military objectives and civilians and civilian objects without distinction, notably because the weapon's effects escape the control of the user and cannot be limited in time and space, as required by IHL.
- The prohibition to carry out disproportionate attacks—i.e. attacks expected to cause incidental civilian casualties and damage to civilian objects that would be excessive in relation to the concrete and direct military advantage anticipated.
- The requirement to take all feasible precautions in attack to avoid or in any event minimize incidental civilian harm. This includes the requirement to do everything feasible to cancel or suspend an attack if it becomes apparent that the

target is not lawful, or that the rule of proportionality could not be respected.

These legal obligations must be fulfilled by those persons who plan, decide on, and carry out an attack, i.e. by the human subjects of IHL, and overall, each of these rules require their human subjects to carry out complex assessments—weighing up complex and not easily quantifiable factors—in the circumstances prevailing at the time of planning and deciding to attack, and also during the attack—i.e. context-based value judgements.

Commanders must make these assessments reasonably proximate in time to the attack. Where these assessments form part of planning assumptions, they must have continuing validity until the execution of the attack. The longer the time-gap between the moment the autonomous weapon is activated by the human operator and the moment the weapon selects and strikes the target, the greater the risk that the assumptions on which the human's IHL judgements are based will no longer be valid, and this is especially so in dynamic (cluttered) environments (e.g. in populated areas). In other words, the facts on the ground may have changed between the moment of activation and the moment the target is struck.

The legal issues raised by the lapse of time between the activation of the weapon and the moment it autonomously selects and strikes the target also comes into play in relation to the IHL requirement that everything feasible be done to cancel or suspend an attack if it becomes apparent that the target is no longer lawful or that the attack may be expected to violate the rule of proportionality.

In sum, to make the context-specific judgments required by IHL rules on the conduct of hostilities, the commander would need to have knowledge of the context, i.e. of the circumstances prevailing at the time of the attack—and in particular of the specific target, its location and surroundings, and the time of the attack. This in turn

demands that the commander have reasonable foreseeability of the weapon's effects when striking the target at the specific time and location. In this respect, commanders rely on the predictability of the weapon and its environment in order to anticipate and limit the weapon's effects as required by IHL—critical predictability for IHL compliance. Unpredictability hinders commanders from properly anticipating and limiting the effects of the weapon as required by IHL. Predictability here is a key—and yet this predictability is eroded by autonomy in weapon systems.

The example of military objectives

To give but one example, consider the definition of a "military objective" under IHL, which in relation to objects requires that attacks be directed only at "military objectives," never against civilian objects. Article 52(2) of Additional Protocol I to the Geneva Conventions defines military objectives as follows:

> military objectives are limited to those objects which by their nature, location, purpose or use make an effective contribution to military action and whose partial or total destruction, capture or neutralization in the circumstances ruling at the time, offers a definite military advantage.

Whether an object makes an "effective contribution to military action," and whether its destruction offers a "definite military advantage," involves weighing up different values, i.e. values-based judgements that cannot readily be reduced to the technical indicators used by machines (autonomous weapons)—i.e. they cannot be reduced to the numerical and quantitative data that are target profiles and information received through sensors and software.

This evaluation is time-bound and, in principle, must be made in relation to a concrete object. With the possible exception of objects

that are by their "nature" military objectives, such as enemy tanks, an object may not be attacked if it does not yet make or no longer makes an effective contribution to the enemy's military action. Sweeping, anticipatory determinations, e.g. declaring all bridges to be military objectives, is not permissible.

Thus, using an autonomous weapon programmed to strike certain objects—e.g. bridges that match its target profile without a contextual evaluation by the human operator of whether the specific bridge that is struck by the weapon makes an effective contribution to military action and whose partial or total destruction in the circumstances ruling at the time, offers a definite military advantage, may fall foul of IHL.

Does IHL already prohibit certain types of AWS ?

Applying the requirements for context-specific judgements and predictability, in the view of the ICRC, AWS that are unsupervised, unpredictable, and unconstrained in time and space would be unlawful under IHL. This would include but is not limited to AWS controlled by AI and machine learning software that is unpredictable or unexplainable.

Key questions left unanswered by existing IHL

But even assuming that everyone agrees with the ICRC's views on the limits that IHL already imposes on the development and use of AWS—which is far from a given—some key questions remain unanswered. These questions stem from the conclusion that existing IHL requires its human subjects to exercise context-dependent value judgments and to have predictability.

The overarching question is what is the level of human control that would allow the commander (the human subject of IHL,

responsible for complying with it) to exercise the context-specific judgments required by IHL?

Specific questions include:

- What is the minimum level of predictability and reliability of the weapon system in its environment of use?
- What constraints are needed for tasks, targets, operational environments, time of operation, and geographical scope of operation?
- What level of human supervision, intervention, and ability to deactivate is needed to comply with IHL rules?

Practical measures of human control to ensure legal compliance and ethical acceptability

As mentioned in the introduction, the ICRC and SIPRI have recently published *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*. Our report recommends three types of control measures, based on humanitarian, legal, ethical, and operational drivers:

Controls on the weapon, such as through limits on: the target-type; the spatial and temporal scope of operation; the weapons' effects; allowing for deactivation and fail-safe mechanisms;

Controls on the environment, such as through situational understanding; excluding protected persons and objects from the area of operation; creating exclusion zones, barriers and warnings; and

Controls through human-machine interaction, such as through human supervision of the system's operation; ability to intervene and deactivate; training the user.

These measures can help reduce or at least compensate for unpredictability inherent in the use of AWS and to mitigate risks.

Let me give some examples for each type of control measure:

Control on the weapon: Constraining the targets and tasks of the autonomous weapon system can help the user of the weapon to predictably limit attacks strictly to military objectives. As an example at air defense systems—weapons that autonomously identify and strike incoming missiles or rockets—their use as we understand it today is typically limited to striking objects with a very specific radar signature, combined with a certain trajectory and velocity. However, this is not a panacea. This must be combined with other constraints or measures to enable a commander to make reasonable assumptions about the AWS's environment of use over the duration of the attack, particularly if the AWS is mobile, or used near a concentration of civilians or civilian objects. This is why, for example, air defense autonomous systems today tend not to be operated in or near civilian airspace/areas. They also tend to be operated under constant human supervision (human on-the-loop).

Control on the environment: Placing constraints on the environment and the context of use, including temporal and spatial limits, may be needed in particular to ensure that the planning assumptions and legal assessments made when activating the AWS remain valid. For example, based on our understanding of how they function, loitering munitions operate in a predetermined geographic area where they search for targets (the "search area"). The loitering weapon is only programmed to "engage" (i.e. detect, select and apply force to) targets such as radar installations while it is within this "search area." If it does not find any targets, it either returns to base or self-destructs.

Control through human-machine interaction: Providing the ability to supervise and intervene in the operation of the AWS during the course of an attack may be needed for all AWS operations, given the inherently dynamic nature of most (if not all) operational environments (however constrained). Commanders would, in most circumstances, need to maintain the ability to supervise the operation

of the weapon and communicate with it after its activation in order to deactivate it if necessary (IHL requires that an attack be cancelled in certain situations).

Our understanding is that, in practice, existing AWS are usually operated within the sorts of constraints described above.

Conclusion

Many critical questions remain open about the limits of what is permissible in terms of AWS use under IHL. In our view, this issue points to the need for internationally agreed limits on AWS, to ensure compliance with IHL and protect humanity. Such limits should build on and strengthen existing IHL and uphold the principles of humanity.

Within the CCW, we see particular value in focusing discussions on Guiding Principles (c) (quality and extent of human-machine interaction required) and (d) (accountability for development and use of AWS) to determine the elements (or criteria) of human control necessary to ensure compliance with international law, including IHL, and ethical acceptability.

There is an urgency to this task due to rapid technological developments that remove or reduce human control over weapons.

International humanitarian law (IHL) and 'LAWS': is there a need for a new protocol?

Rio Seminar on Autonomous Weapons Systems

20 February 2020

Kathleen Lawand
Head, Arms Unit
Legal Division

ICRC

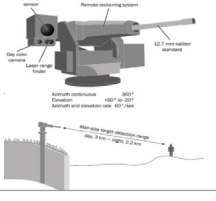
Introduction

- **Autonomy in weapon systems**
  - Weapon selects and attacks target without human intervention
  - Loss of human control over the use of force
- **ICRC's core considerations**
  - Humanitarian consequences
  - Compatibility with IHL
- **Three key points**
  1. Protection is afforded by existing IHL
  2. But also significant challenges for IHL compliance
  3. Critical questions for legal compliance and ethical acceptability must be urgently addressed

ICRC

Protection afforded by existing IHL

- **IHL rules on conduct of hostilities**
  - Obligation to distinguish military objectives from civilians/civilian objects
  - Prohibition of indiscriminate and disproportionate attacks
  - Obligation to take all feasible precautions in attack, to avoid or in any event minimize civilian harm
- **IHL requires the (human) commander / combatant to**
  - make complex, context-specific value judgements
  - based on the circumstances prevailing at the time of attack
  - predict the consequences of the attack



ICRC

Challenges in complying with IHL

- The **unique characteristics of autonomous weapons** -- loss of human control and unpredictability in the consequences of their use -- present **unique challenges** to complying with IHL and raise profound ethical concerns

1. **Context-specific value judgements**
   - the characterization of the target as lawful or unlawful generally involves qualitative judgements
   - the value judgements of the commander cannot be reduced to the technical indicators (numerical and quantitative data) used by machines
   - for example: IHL definitions of “military objectives” and “proportionality in attack”, which require a weighing up of different values

ICRC

Challenges in complying with IHL

Military objectives: "those objects which by their nature, location, purpose or use make an effective contribution to military action and whose partial or total destruction, capture or neutralization in the circumstances ruling at the time, offers a definite military advantage"

Proportionality: prohibition to conduct an "attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated"

- context-based value judgements by those (humans) who plan, decide upon and carry out attacks
- weighing up different qualitative values that change over time

ICRC

Challenges in complying with IHL

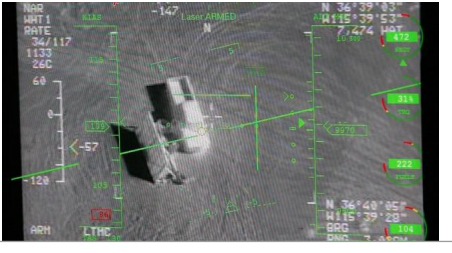
2. Unpredictability

- Autonomous weapons raise concerns about unpredictability, as it is the weapon itself that selects a specific target and the time and location of attack
- Unpredictability of the weapon and of the environment hinders the commander from properly anticipating and limiting the weapon's effects, as required by IHL
- Assumption on which the commander plans and decides to attack must remain valid until the execution of the attack
  - Facts on the ground may have changed between weapon's activation and the moment it selects and attacks the target

ICRC

Control measures

- **Three types of control measures to ensure AWS can be used in compliance with IHL:**
  - constrain the weapon's targets and tasks
  - constrain the environment and situation of use, including through temporal and spatial limits
  - retain the ability to supervise and intervene in the operation of AWS during the course of an attack



ICRC

Critical questions for IHL compliance and ethical acceptability

- **Build on broad agreement on "human control"**
  - Determine which elements of human control are needed to ensure compliance with IHL and ethical acceptability
- **Critical questions:**
  - Is it legally and ethically acceptable to develop and use autonomous weapons designed to use force against persons?
  - And against objects in areas where civilians and civilian objects are at risk?
  - What limits should be set on the use of autonomous weapons to address their unpredictability?
    - Limits on tasks, duration (time-frame) and area (geographical scope)?

ICRC

ICRC

Rio Seminar on Autonomous Weapon Systems
Naval War College, Rio de Janeiro
19-20 February 2020

AUTONOMOUS WEAPON SYSTEMS[1]

The ICRC understands autonomous weapon systems as: *Any weapon system with autonomy in its critical functions. That is, a weapon system that can select and attack targets without human intervention*. Autonomy in critical functions – already found in some existing weapons to a limited extent, such as air defence systems, active protection systems, and some loitering weapons – is a feature that could be incorporated in any weapon system.

The most important aspect of autonomy in weapon systems – from a humanitarian, legal and ethical perspective – is that the weapon system self-initiates, or triggers, an attack in response to its environment, based on a generalized target profile. To varying degrees, the user of the weapon will know neither the specific target nor the exact timing and location of the attack that will result. Autonomous weapon systems are, therefore, clearly distinguishable from other weapon systems, where the specific timing, location and target are chosen by the user at the point of launch or activation.

The ICRC's primary concern is loss of human control over the use of force as a result of autonomy in the critical functions of weapon systems. Depending on the constraints under which a system operates, the user's uncertainty about the exact timing, location and circumstances of the attack(s) may put civilians at risk from the unpredictable consequences of the attack(s). It also raises legal questions, since combatants must make context specific judgements to comply with IHL. And it raises ethical concerns as well, because human agency in decisions to use force is necessary in order to uphold moral responsibility and human dignity.

Fuller understanding of the legal,[2] military,[3] ethical,[4] and technical[5] aspects of autonomous weapon systems has enabled the ICRC to refine its views.[6] It continues to espouse a human-centred approach, based on its reading of the law and ethical considerations for humans in armed conflict.[7]

Human control under IHL

The ICRC holds that legal obligations under IHL rules on the conduct of hostilities must be fulfilled by those persons who plan, decide on, and carry out military operations. It is humans, not machines, that comply with and implement these rules, and it is humans who can be held accountable for violations. Whatever the

[1] Extract from ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts,* 2019 pp 29-31; available at https://www.icrc.org/en/document/icrc-report-ihl-and-challenges-contemporary-armed-conflicts
[2] Neil Davison, "A legal perspective: Autonomous weapon systems under international humanitarian law", in UNODA Occasional Papers, No. 30, November 2017; available at https://www.icrc.org/en/document/autonomous-weaponsystems-under-international-humanitarian-law; ICRC, *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects,* 2014: available at https://www.icrc.org/en/document/report-icrc-meeting-autonomousweapon-systems-26-28-march-2014.
[3] See ICRC, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons*, 2016; available at https://www.icrc.org/en/publication/4283-autonomous-weapons-systems.
[4] See ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?,* 2018; available at https://www.icrc.org/en/document/ethics-and-autonomous-weapon-systems-ethical-basis-human-control.
[5] See ICRC, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control*, 2019; available at https://www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.
[6] See ICRC, *IHL Challenges Report* 2011, pp. 39–40. On definitions in particular, see ICRC, IHL Challenges Report 2015, p. 45.
[7] See ICRC, Statements to the Group of Governmental Experts on Lethal Autonomous Weapons Systems, March 2019; available at https://www.unog.ch/80256EE600585943/(httpPages)/5535B644C2AE8F28C1258433002BBF14?OpenDocument.

machine, computer program, or weapon system used, individuals and parties to conflicts remain responsible for their effects.

Certain limits on autonomy in weapon systems can be deduced from existing rules on the conduct of hostilities – notably the rules of distinction, proportionality and precautions in attack – which require complex assessments based on the circumstances prevailing at the time of the decision to attack, but also during an attack. Combatants must make these assessments reasonably proximate in time to the attack. Where these assessments form part of planning assumptions, they must have continuing validity until the execution of the attack. Hence, commanders or operators must retain a level of human control over weapon systems sufficient to allow them to make context-specific judgments to apply the law in carrying out attacks.

Human control can take various forms during the development and testing of a weapon system ("development stage"); the taking of the decision to activate the weapon system ("activation stage"); and the operation of the weapon system as it selects and attacks targets ("operation stage"). Human control at the activation and operation stages is the most important factor for ensuring compliance with the rules on the conduct of hostilities. Human control during the development stage provides a means to set and test control measures that will ensure human control in use. However, control measures at the development stage alone – meaning control in design – will not be sufficient.

Importantly, however, existing IHL rules do not provide all the answers. Although States agree on the importance of human control – or "human responsibility"[8] – for legal compliance, opinion varies on what this means in practice. Further, purely legal interpretations do not accommodate the ethical concerns raised by the loss of human control over the use of force in armed conflict.

Towards limits on autonomy in weapon systems

In the ICRC's view, the unique characteristics of autonomous weapon systems, and the associated risks of loss of control over the use of force in armed conflict, mean that internationally agreed limits are needed to ensure compliance with IHL and to protect humanity.

Insofar as the sufficiency of existing law – particularly IHL – is concerned, it is clear, as shown above, that existing IHL rules – in particular distinction, proportionality, and precautions in attack – already provide limits to autonomy in weapon systems. A weapon with autonomy in its critical functions that is unsupervised, unpredictable and unconstrained in time and space would be unlawful, because humans must make the context-specific judgments that take into account complex and not easily quantifiable rules and principles.

However, it is also clear that existing IHL rules do not provide all the answers. What level of human supervision, intervention and ability to deactivate is needed? What is the minimum level of predictability and reliability of the weapon system in its environment of use? What constraints are needed for tasks, targets, operational environments, time of operation, and geographical scope of operation?

Moreover, the limits dictated by ethical concerns may go beyond those found in existing law. Anxieties about the loss of human agency in decisions to use force, diffusion of moral responsibility, and loss of human dignity are most acute with autonomous weapon systems that present risks for human life, and especially with the notion of anti-personnel systems designed to target humans directly. The principles of humanity may demand limits on or prohibitions against particular types of autonomous weapon and/or their use in certain environments.

At a minimum, there remains an urgent need for agreement on the type and degree of human control necessary in practice to ensure both compliance with IHL and ethical acceptability.

[8] United Nations, Report of the 2018 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2018/3, 23 October 2018.

Challenges Towards a Regulatory Framework

Ambassador Michael Biontino
Senior Special Adviser from the German Federal Foreign Office

First of all, I would like to thank the Brazilian Ministry of Foreign Affairs, FUNAG, and the Brazilian Naval War College (EGN) for having invited me to participate in this seminar. I feel very much honored and will endeavor to contribute to the discussion on Autonomous Weapons Systems from the angle of International Law and IHL.

We have been asked if there is a **need for a new protocol** under the CCW: The CCW seeks to prohibit or restrict the use of weapons which may be deemed excessively injurious or whose effects are indiscriminate. Given that the views range from, on the one side, that current IHL is quite sufficient to deal with LAWS as well, to, on the other side, that there is an urgent need for a comprehensive legally binding instrument to prohibit LAWS, each view being substantiated

by serious arguments, it seems that this is very much a political issue, where a strictly legal analysis can only be of limited value.

I suggest, therefore, that we take as a **starting point the CCW's discussion as it stands right now**.

The CCW, in its 2019 meeting, decided that the task of the GGE in 2020 would essentially be the clarification, consideration, and development of aspects of the **normative and operational framework** on emerging technologies in the area of Lethal Autonomous Weapons Systems. This, indeed, suggests a regulatory framework for LAWS, whose details need to be elaborated, based, of course, on the GGE's conclusion that IHL applies fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems.

In this context, I would point out that the German government has clearly positioned itself by advocating for LAWS understood as lethal weapons deprived of human control to be renounced—or morally condemned—globally. This does not mean that the German government holds that future systems characterized by a high degree of autonomy should be generally prohibited.

For our discussion to be able to make a meaningful contribution to the work of the GGE in 2020, I believe we have to **identify the challenges** it will face. In this vein a further analysis of the normative elements of a regulatory framework, whatever its nature may be, seems appropriate. These elements are, in particular, scope of application, general obligations, definition, verification, and possibly transparency and confidence building.

Scope of application

CCW Protocol V and Amended Protocol II limit their scope of application to, generally speaking, **terrestrial scenarios**[1] (i.e. ground-based weapons systems). This limited scope is indeed appropriate, given the intrinsic nature of weapons like mines, booby-traps or explosive remnants of war and the humanitarian consequences associated with their use. Other scenarios would have been of no or very limited relevance.

Concerning LAWS, however, a **wider scope seems to be relevant**.[2] Beyond ground-to ground scenarios, in particular air-to-ground, ground-to-air, air-to-air, maritime, cyber-, and possibly outer space scenarios would have to be considered, since, indeed, the potential military utility for deploying them—e.g. speed of decision making and/or lack of reliable communication—might be clearer in these other scenarios.

This has far-reaching consequences concerning requirements, namely in terms of distinction, proportionality, and precaution in attack that LAWS will have to fulfill in order to be compliant with IHL.

Whereas in ground-to-ground scenarios, particularly in a **cluttered environment**, for LAWS to be able to distinguish autonomously legitimate targets and evaluate if the military value of this target justifies a certain collateral damage, and to take the decision to engage these targets seems, under present circumstances,

1 CCW Protocol V, Art 1; Para. 2: "This Protocol shall apply to explosive remnants of war on the land territory including internal waters of High Contracting Parties".
CCW amended Protocol II: "This Protocol relates to the use on land of the mines, booby-traps and other devices, defined herein, including mines laid to interdict beaches, waterway crossings or river crossings, but does not apply to the use of anti-ship mines at sea or in inland waterways.

2 The CCW GGE Report 2019 (CCW/GGE.1/2019/3) used the terms "operational context" and concluded that "further work is needed to build shared understanding on the role of operational constraints regarding tasks, target profiles, time-frame of operation, and scope of movement over an area and operating environment" without going into further detail.

technically quite an impossible task. However—it is clear—autonomy would have to be very carefully analyzed throughout the entire targeting and engagement cycle, ranging from identifying the military objectives to the critical functions of target selection and engagement.

On the other hand, in **air-to-air and maritime scenarios**, in order to identify legitimate targets, LAWS could, as technology stands right now, possibly rely on a database of pre-identified targets.[3] Furthermore, the issue of collateral damage would not be relevant to the same degree. This implies that the requirements of distinction, proportionality, and precaution under IHL would possibly be a realistic aspiration.

Future negotiations for a regulatory framework for LAWS might, therefore, have to face the challenge of clearly distinguishing different scenarios in defining the scope of application and the appropriate level of obligations, i.e. pre- or proscriptive measures, for each scenario. This, of course, within the regulatory framework of IHL prohibiting indiscriminate attacks and excessive collateral damage, which has to be complied with in any scenario of the use of LAWS.

General obligations

The CCW and its protocols, in principle, prohibit the **use of certain weapons**. However, the GGE's discussion has gone beyond simply the "use" of LAWS. The question has been raised of whether a comprehensive approach should be followed, including the development, production, deployment, acquisition, use, and transfer of LAWS. Such a broader approach implies substantive challenges for future negotiations.

3 Notwithstanding the fact that this might question if they would be fully autonomous systems.

It is understood that autonomous systems are being developed in the civilian sphere, and a majority of sub-components of any LAWS would be of a **dual-use nature**. A regulatory framework on LAWS, as a consequence, would have to ensure that relevant obligations do not infringe on progress in or access to legitimate civilian research and developments.

In a second step, it seems useful to differentiate between the **different scenarios** already mentioned above (i.e. ground-to-ground, air-to-ground, ground-to-air, air-to-air, maritime, and possibly outer space scenarios) in order to specify the obligations under IHL governing the development and the use of emerging technologies in the area of LAWS. A "one size fits all" approach seriously risks being dysfunctional, for instance, when comparing a highly complex urban setting with a submarine or outer space scenario.

When it comes to development, production, deployment, acquisition, use, and transfer,[4] the following options[5] have been discussed in the CCW GGEs:

- one option discussed is a set of **guidelines and/or positive commitmentsfor future developments for LAWS**. The "11 Guiding Principles," as endorsed by the CCW in its 2019 meeting, contain such an approach, which, in particular, could reconfirm important principles such as the necessity of accountability to be ensured, including through a responsible chain of human command and control. However, in order to be able to serve as guidelines and and/or positive commitments for future developments, the **"Guiding Principles" would**

4 In this context, proliferation issues and the risks of misuse by non-state actors have been highlighted.

5 In this context, it should be recalled that some GGE participants held the view that no further legal measures were needed, if the view that IHL is fully applicable and sufficient to deal with any possible challenges raised by LAWS is considered.

probably have to be operationalized in a way so that concrete conclusions could be drawn.

- A regulatory framework containing such positive commitments could be effective, naturally with different degrees of stringency, irrespective of its legal nature, ranging from a politically binding declaration, a Code of Conduct, to a politically or legally binding instrument.
- Another option discussed is a regulatory **framework containing prohibitions** in the development, production, deployment, acquisition, use, and transfer of LAWS. In this case, it was argued that only a politically or legally binding instrument would guarantee the required amount of stringency.
- As a third option, a ***moratorium*** on the development and use of LAWS in the interim has been proposed.

As future negotiations discuss these options in more detail, they will, in particular, have to face the challenge of:

- which option is the most appropriate in the different scenarios; and
- to what extent effectiveness will depend on a shared understanding/definition of what constitutes a "lethal autonomous weapons system".

Definition

Definitions have always been a central element for any regulatory framework in arms control and disarmament. However, approaches have varied widely. Whereas some regimes in the nuclear area, for instance the NPT (Treaty on the Non-Proliferation of Nuclear Weapons) and the TPNW (Treaty on the Prohibition of Nuclear Weapons) are based on a general understanding of what a nuclear weapon is, on the other side, in particular in areas where dual-use issues are very important, the Chemical Weapons Convention, for

instance, follows an inverted approach in defining what a chemical weapon is, prohibiting all toxic chemicals "except for purposes not prohibited by the Chemical Weapons Convention". The key issue always has been to **clearly delimit what systems are to be covered** by the obligations under the regulatory framework and to ensure that no relevant systems are left out.[1]

As the discussion in the past CCW GGEs has shown, the definition of what a LAWS is remains a major challenge, in particular to distinguish clearly between autonomous and automated or remote-controlled weapon systems, that already exist.

Furthermore, it has become evident that any regulatory framework would ideally have to clearly distinguish LAWS from other systems, in particular automatic systems with predefined targets.

There is a general understanding that **autonomy in weapons systems will evolve gradually**. The question, therefore, remains: autonomy in which functions? Here again a critical analysis of autonomy, i.e. the degree of human-machine interactions at various stages of the life cycle of a weapons system, including the targeting and engagement cycle, seems to be required.

In the GGE's discussion, it has become apparent that **technical approach** to autonomy, such as physical performance, endurance or sophistication in targeting acquisition and engagement may alone not be sufficient to characterize lethal autonomous weapons systems, especially in view of:

- rapid evolution in technology; and
- the fact that autonomy may be viewed as a spectrum, with difficulties delineating between such concepts of automation,

1 The Convention on Cluster Munitions has been, for instance, questioned for not covering all relevant systems.

semi- or fully-autonomous—and that the term covers a wide range of technical capabilities.

However, in 2019 the GGE concluded that **human judgement is essential** in order to ensure that the potential use of weapons systems based on emerging technologies in the area of Lethal Autonomous Weapons Systems is in compliance with international law, and in particular IHL.[2] As a consequence, the issue arises of to what extent and in which part of the life cycle of a weapons system, including the targeting and engagement cycle, human judgement is to be retained.

A regulatory framework on LAWS should in particular address systems which operate beneath this threshold of required human judgement.

Such a "**normative approach**" to a definition on LAWS seems, indeed, the most promising one. However, the challenge for future negotiations will be to operationalize the required degree of human judgement to serve a definition.

In summary, the discussion on definitions will set the stage for what commitments could be envisaged for a regulatory framework on LAWS, since the **granularity** required depends very much on the substantive nature of obligations and their legal framing:

- on the one hand, for a set of guidelines for future developments for LAWS contained in a **politically binding** declaration or a Code of Conduct, a general characterization or understanding of LAWS might be sufficient; but

2 See in particular Principle 11: Human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon, should ensure that the potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems is in compliance with applicable international law, in particular International Humanitarian Law (IHL).

- for a verifiable politically or even **legally binding instrument** containing more stringent obligations, going up to a comprehensive prohibition, clear-cut definitions would be indispensable.

Verification, transparency, and confidence building

Given the important strategic, military, and international security dimensions of LAWS, verification and transparency will be of paramount importance.

Let me, in this context, recall the **functions of verification and transparency**. It is generally understood that verification should:

create confidence among the relevant actors that arms control and disarmament commitments are adhered to;

detect a militarily significant violation of the underlying arms control or disarmament agreement in time; and

ensure that appropriate counter action in case of non-compliance can be taken, *inter alia*, to enable to respond effectively and possibly to deny the violator the benefits of the violation.

The question is, indeed, **how to achieve verification or transparency** in an effective and efficient manner. When discussing these challenges in detail, security and proprietary concerns will have to be respected appropriately.

However, they could **build on precedence**, which in my view could be usefully combined:

Article 36 of the Additional Protocol I to the Geneva Conventions on national legal weapons reviews: One of the key concerns in dealing with LAWS in the CCW or the wider IHL/HumanRights context is whether these systems can predictably and reliably[3] be

3 Since autonomous decision-making, self-learning and the associated unpredictability, and possibly the ability to redefine missions or objectives independently are assumed to be an integral part of

used in conformity with IHL, i.e. do they comply, in particular, with the requirements of distinction, proportionality and precaution in attack? This is precisely the purview of Art. 36 procedures.

We have to realize, however, that the **issue of verification, transparency, and confidence building has barely been addressed**. As a consequence, work on appropriate verification techniques and procedures would have to be initiated. Scientific input would be most valuable in this context.

Concluding remarks

The **elements of a regulatory framework**, such as "scope of application, definition, general obligations, and possibly verification and transparency and confidence building" depend on and are interrelated with each other. It is hardly possible to address them in isolation, as the GGE in 2020 will:

- consider the legal, technological and military aspects, as well as the interaction between them, bearing in mind ethical considerations; and
- with a perspective towards a normative and operational framework on emerging technologies in the area of lethal autonomous weapons systems.

Given that the CCW GGEs on LAWS have not discussed all these elements in the appropriate detail, the following pragmatic approach could be considered:

- Start with clearly identifying **Guiding Principles that may need further operationalization**. In this area progress seems possible, in particular, since it does not have to be based on a clear-cut definition;

LAWS, this will probably pose novel challenges to weapons review processes. Assuranceswould be required that their employment will predictably and reliably be in conformity with IHL.

- Determine the **appropriate framework** for such guidelines, i.e. a politically binding declaration or instrument, or a Code of Conduct;
- Start work on the various scenarios and challenges for IHL conformity;
- Work on an operational definition for LAWS, and, in parallel, start work on verification, transparency and confidence building;
- To culminate with different requirements or standards for the precautionary measures needed in different scenarios.

The Need for and Elements of a New Treaty on Fully Autonomous Weapons

Bonnie Docherty[4]
Harvard Law School

The rapid evolution of autonomous technology threatens to strip humans of their traditional role in the use of force. Fully autonomous weapons, in particular, would select and engage targets without meaningful human control. Due in large part to their lack of human control, these systems, also known as LAWS or "killer robots," raise a host of legal and ethical concerns.

4 Bonnie Docherty is a lecturer on law and the Associate Director of Armed Conflict and Civilian Protection at Harvard Law School's International Human Rights Clinic. She is also a senior researcher in the Arms Division of Human Rights Watch, which has coordinated the Campaign to Stop Killer Robots since its inception in 2012. Docherty has participated in every Convention on Conventional Weapons meeting about lethal autonomous weapons systems and has published extensively on the topic. See Human Rights Watch and the Harvard Law School International Human Rights Clinic (IHRC), "Reviewing the Record: Reports on Killer Robots," <http://hrp.law.harvard.edu/wp-content/uploads/2018/08/Killer_Robots_Handout.pdf> (accessed May 22, 2020).

States parties to the CCW have held eight in-depth meetings on lethal autonomous weapons systems since 2014. They have examined the extensive challenges raised by the systems and recognized the importance of retaining human control over the use of force. Progress toward an appropriate multilateral solution, however, has been slow. If states do not shift soon from abstract talk to treaty negotiations, the development of technology will outpace international diplomacy.

Approaching the topic from a legal perspective, this chapter argues that fully autonomous weapons cross the threshold of acceptability and should be banned by a new international treaty. The chapter first examines the concerns raised by fully autonomous weapons, particularly under International Humanitarian Law. It then explains why a legally binding instrument best addresses those concerns. Finally, it proposes key elements of a new treaty to maintain meaningful human control over the use of force and prohibit weapons systems that operate without it.

The problems posed by fully autonomous weapons

Fully autonomous weapons would present significant hurdles to compliance with International Humanitarian Law's fundamental rules of distinction and proportionality.[5] In today's armed conflicts, combatants often seek to blend in with the civilian population. They hide in civilian areas and wear civilian clothes. As a result, the ability to distinguish combatants from civilians or those *hors de combat* often requires gauging an individual's intentions based on subtle behavioral cues, such as body language, gestures, and tone of voice. Humans, who can relate to other people, can better interpret those cues than inanimate machines.[6]

5 Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), adopted June 8, 1977, 1125 U.N.T.S. 3, entered into force December 7, 1978, arts. 48 and 51(4-5).

6 Human Rights Watch and IHRC, Making the Case: The Dangers of Killer Robots and the Need for a

Fully autonomous weapons would find it even more difficult to weigh the proportionality of an attack. The proportionality test requires determining whether expected civilian harm outweighs anticipated military advantage on a case-by-case basis in a rapidly changing environment. Evaluating the proportionality of an attack involves more than a quantitative calculation. Commanders apply human judgment, informed by legal and moral norms and personal experience, to the specific situation. Whether the human judgment necessary to assess proportionality could ever be replicated in a machine is doubtful. Furthermore, robots could not be programmed in advance to deal with the infinite number of unexpected situations they might encounter on the battlefield.[7]

The use of fully autonomous weapons also risks creating a serious accountability gap.[8] International Humanitarian Law requires that individuals be held legally responsible for war crimes and grave breaches of the Geneva Conventions. Military commanders or operators could be found guilty if they deployed a fully autonomous weapon with the intent to commit a crime. It would, however, be legally challenging and arguably unfair to hold an operator responsible for the unforeseeable actions of an autonomous robot.

Finally, fully autonomous weapons contravene the Martens Clause, a provision that appears in numerous International Humanitarian Law treaties.[9] The clause states that if there is no specific law on a topic, civilians are still protected by the principles

Preemptive Ban (December 2016), <https://www.hrw.org/sites/default/files/report_pdf/arms1216_web.pdf> (accessed May 21, 2020), p. 5.

7 Ibid., pp. 5-8.

8 See generally Human Rights Watch and IHRC, ***Mind the Gap: The Lack of Accountability for Killer Robots*** (April 2015), <https://www.hrw.org/sites/default/files/reports/arms0415_ForUpload_0.pdf> (accessed May 20, 2020).

9 See generally Human Rights Watch and IHRC, ***Heed the Call: A Moral and Legal Imperative to Ban Killer Robots*** (August 2018), <https://www.hrw.org/sites/default/files/report_pdf/arms0818_web.pdf> (accessed May 20, 2020).

of humanity and dictates of public conscience.[10] Fully autonomous weapons would undermine the principles of humanity because of their inability to show compassion or respect human dignity.[11] Widespread opposition to fully autonomous weapons among faith leaders, scientists, tech workers, civil society organizations, the public, and more indicates that this emerging technology also runs counter to the dictates of public conscience.[12]

Fully autonomous weapons pose numerous other threats that go far beyond concerns over compliance with International Humanitarian Law. For many, delegating life-and-death decisions to machines would cross a moral red line.[13] The use of fully autonomous weapons, including in law enforcement operations, would undermine the rights to life, remedy, and dignity.[14] Development and production

10 See, for example, Convention (II) with Respect to the Laws and Customs of War on Land and its Annex: Regulations concerning the Laws and Customs of War on Land, The Hague, adopted July 29, 1899, entered into force September 4, 1900, pmbl., para. 8; Protocol I, art. 1(2).

11 Human Rights Watch and IHRC, ***Heed the Call***, pp. 19-27.

12 See, for example, PAX, "Religious Leaders Call for a Ban on Killer Robots," November 12, 2014, <https://www.paxforpeace.nl/stay-informed/news/religious-leaders-call-for-a-ban-on-killer-robots>; "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," opened for signature July 28, 2015, <https://futureoflife.org/open-letter-autonomous-weapons/?cn-reloaded=1> (signed, as of May 2020, by 4,502 AI and robotics researchers and 26,215 others); Scott Shane and Daisuke Wakabayashi, "'The Business of War': Google Employees Protest Work for the Pentagon," ***New York Times***, April 4, 2018, <https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html?partner=IFTTT>; Campaign to Stop Killer Robots, "Learn: The Threat of Fully Autonomous Weapons," <https://www.stopkillerrobots.org/learn/>; Ipsos, "Six in Ten (61%) Respondents across 26 Countries Oppose the Use of Lethal Autonomous Weapons Systems," January 21, 2019, <https://www.ipsos.com/en-us/news-polls/human-rights-watch-six-in-ten-oppose-autonomous-weapons> (all accessed May 21, 2020). See also Human Rights Watch and IHRC, ***Heed the Call***, pp. 28-43.

13 UN Human Rights Council, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns, "Lethal Autonomous Robotics," <http://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A-HRC-23-47_en.pdf> (accessed May 21, 2020), p. 17 (writing, "Machines lack morality and mortality, and should as a result not have life and death powers over humans").

14 See generally Human Rights Watch and IHRC, ***Shaking the Foundations: The Human Rights Implications of Killer Robots*** (May 2014), <https://www.hrw.org/sites/default/files/reports/arms0514_ForUpload_0.pdf> (accessed May 20, 2020). See also Heyns, "Lethal Autonomous Robotics," pp. 6 (on the right to life: "the introduction of such powerful yet controversial new weapons systems has the potential to pose new threats to the right to life"), 15 (on the right to

of these machines could trigger an arms race, and the systems could proliferate to irresponsible states and non-state armed groups.[15] Even if new technology could address some of the International Humanitarian Law problems discussed above, it would not resolve many of these other concerns.

The need for a legally binding instrument

The unacceptable risks posed by fully autonomous weapons necessitate creation of a new legally binding instrument. It could take the form of a stand-alone treaty or a protocol to the Convention on Conventional Weapons. Existing international law, including International Humanitarian Law, is insufficient in this context because its fundamental rules were designed to be implemented by humans, not machines. At the time states negotiated the additional protocols to the Geneva Conventions, they could not have envisioned full autonomy in technology. Therefore, while CCW states parties have agreed that international humanitarian law applies to this new technology, there are debates about how it does.[16]

A new treaty would clarify and strengthen existing international humanitarian law. It would establish clear international rules to address the specific problem of weapons systems that operate outside of meaningful human control. In so doing, the instrument would fill the legal gap highlighted by the Martens Clause, help eliminate

remedy: "If the nature of a weapon renders responsibility for its consequences impossible, its use should be considered unethical and unlawful as an abhorrent weapon"), and 20 (on dignity: "there is widespread concern that allowing [fully autonomous weapons] to kill people may denigrate the value of life itself").

15 "Autonomous Weapons: An Open Letter from AI & Robotics Researchers"; Human Rights Watch and IHRC, ***Making the Case***, pp. 29-30.

16 The applicability of International Humanitarian Law to lethal autonomous weapons systems is the first of 11 guiding principles adopted by CCW states parties. "Report of the 2018 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems," CCW/GGE.1/2018/3, October 23, 2018, <https://www.unog.ch/80256EDD006B8954/(httpAssets)/20092911F6495FA7C125830E003F9A5B/$file/CCW_GGE.1_2018_3_final.pdf> (accessed May 21, 2020), para. 26(a).

disputes about interpretation, promote consistency of interpretation and implementation, and facilitate compliance and enforcement.[17]

The treaty could also go beyond the scope of current international humanitarian law. While the relevant provisions of International Humanitarian Law focus on the use of weapons, a new treaty could address development, production, and use. In addition, it could apply to the use of fully autonomous weapons in both law enforcement operations as well as situations of armed conflict.[18]

A legally binding instrument is preferable to the "normative and operational framework" that the CCW states parties agreed to develop in 2020 and 2021.[19] The phrase "normative and operational framework" is intentionally vague, and thus has created uncertainty about what states should be working toward. While the term could encompass a legally binding CCW protocol, it could also refer to political commitments or voluntary best practices, which would not be enough to preempt what has been called the "third revolution in warfare."[20] Whether adopted under the auspices of CCW or in another forum, a legally binding instrument would bind states parties to clear obligations. Past experience shows that the stigma it would create could also influence states not party and non-state armed groups.

The elements of a new treaty

CCW states parties have discussed the problems of fully autonomous weapons and the adequacy of International Humanitarian Law since 2014. It is now time to move forward and determine the specifics of an effective response. This chapter will

17 Campaign to Stop Killer Robots, "Key Elements of a Treaty on Fully Autonomous Weapons: Frequently Asked Questions," February 2020, <https://www.stopkillerrobots.org/wp-content/uploads/2020/03/FAQ-Treaty-Elements.pdf> (accessed May 21, 2020), p. 2.

18 Ibid.

19 CCW Meeting of High Contracting Parties, "Final Report," CCW/MSP/2019/9, December 13, 2019, <https://undocs.org/CCW/MSP/2019/9> (accessed May 21, 2020), para. 31.

20 "Autonomous Weapons: An Open Letter from AI & Robotics Researchers."

lay out key elements of a proposed treaty, which were drafted by the International Human Rights Clinic at Harvard Law School and adopted by the Campaign to Stop Killer Robots in 2019.[1]

The proposal outlined below does not constitute specific treaty language. States will determine the details of content and language over the course of formal negotiations. Instead, the proposal highlights elements that a final treaty should contain in order to effectively address concerns that many states, international organizations, and civil society have identified. The elements include the treaty's scope, the underlying concept of meaningful human control, and core obligations.

Scope

The proposal for a new treaty recommends a broad scope of application. The treaty should apply to any weapon system that selects and engages targets based on sensor processing, rather than human input.[2] The breadth of scope aims to ensure that all systems in that category—whether current or future—are assessed, and that problematic systems do not escape regulation. The prohibitions and restrictions, which are detailed below, however, are future-looking and focus on fully autonomous weapons.

Meaningful human control

The concept of meaningful human control is crucial to the new treaty because the moral, legal, and accountability problems

1 Campaign to Stop Killer Robots, "Key Elements of a Treaty on Fully Autonomous Weapons," November 2019, <https://www.stopkillerrobots.org/wp-content/uploads/2020/04/Key-Elements-of-a-Treaty-on-Fully-Autonomous-WeaponsvAccessible.pdf> (accessed May 21, 2020). See also Campaign to Stop Killer Robots, "Key Elements of a Treaty on Fully Autonomous Weapons: Frequently Asked Questions."

2 Article 36, "Autonomy in Weapons Systems: Mapping a Structure for Regulation through Specific Policy Questions," November 2019, <http://www.article36.org/wp-content/uploads/2019/11/regulation-structure.pdf> (accessed May 21, 2020), p. 1.

associated with fully autonomous weapons are largely attributable to the lack of such control.[3] Recognizing these risks, most states have embraced the principle that humans must play a role in the use of force.[4] While they have used different terminology, many states and experts prefer the term "meaningful human control." "Control" is stronger than alternatives such as "intervention" and "judgment" and is broad enough to encompass both of them; it is also a familiar concept in international law.[5] "Meaningful" ensures that control rises to a significant level.[6]

States, international organizations, non-governmental organizations, and independent experts have identified numerous components of meaningful human control.[7] This chapter distills those components into three categories:

- Decision-making components give humans the information and ability to make decisions about whether the use of

3 Human Rights Watch and IHRC, "Killer Robots and the Concept of Meaningful Human Control," April 2016, <https://www.hrw.org/sites/default/files/supporting_resources/robots_meaningful_human_control_final.pdf> (accessed May 21, 2020), pp. 2-6.

4 Ray Acheson, "It's Time to Exercise Human Control over the CCW," Reaching Critical Will's ***CCW Report***, vol. 7, no. 2, March 27, 2019, <https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2019/gge/reports/CCWR7.2.pdf> (accessed May 21, 2020), p. 2 (reporting that "[o]nce discussions got under way, it became clear that the majority of governments still agree human control is necessary over critical functions of weapon systems").

5 Campaign to Stop Killer Robots, "Key Elements of a Treaty on Fully Autonomous Weapons: Frequently Asked Questions," p. 5.

6 According to Article 36, "The term 'meaningful' can be argued to be preferable because it is broad, it is general rather than context specific (e.g. appropriate) [and] derives from an overarching principle rather being outcome driven (e.g. effective, sufficient)." Article 36, "Key Elements of Meaningful Human Control," April 2016, <http://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf> (accessed May 21, 2020), p. 2.

7 See, for example, Allison Pytlak and Katrin Geyer, "News in Brief," Reaching Critical Will's ***CCW Report***, vol. 7, no. 2, March 27, 2019, pp. 10-12 (summarizing states' views on control from one CCW session); International Committee of the Red Cross, "Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control," August 2019, <https://www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control>; International Committee for Robot Arms Control, "What Makes Human Control over Weapons Systems 'Meaningful'?" August 2019, <https://www.icrac.net/wp-content/uploads/2019/08/Amoroso-Tamburrini_Human-Control_ICRAC-WP4.pdf>; iPRAW, Focus on Human Control (August 2019), <https://www.ipraw.org/wp-content/uploads/2019/08/2019-08-09_iPRAW_HumanControl.pdf> (all accessed May 21, 2020).

force complies with law and ethics. For example, a human operator should have: an understanding of the operational environment; an understanding of how the system functions, such as what it might identify as target; and sufficient time for deliberation;

- Technological components are embedded features of a weapon system that enhance meaningful human control. Technological components include, for example, predictability and reliability, the ability of the system to relay information to a human operator, and the ability of a human to intervene after activation of the system; and
- Operational components limit when and where a weapon system can operate and what it can target. Factors that could be constrained include the time between a human's legal assessment and a system's application of force, the duration of a system's operation, and the nature and size of the geographic area of operation.[8]

None of these components are independently sufficient, but they each increase the meaningfulness of control, and they often work in tandem. The above list may not be exhaustive; further analysis of existing and emerging technologies may reveal others. Regardless, a new legally binding instrument should incorporate such components as prerequisites for meaningful human control.

Core obligations

The heart of a legally binding instrument on fully autonomous weapons should consist of a general obligation combined with prohibitions and positive obligations.[9]

8 Campaign to Stop Killer Robots, "Key Elements of a Treaty on Fully Autonomous Weapons," pp. 3-4.

9 These obligations are drawn from the Campaign to Stop Killer Robots, "Key Elements of a Treaty on Fully Autonomous Weapons."

General obligation

The treaty should include a general obligation for states to maintain meaningful human control over the use of force. This obligation establishes a principle to guide interpretation of the rest of the treaty. Its generality is designed to avoid loopholes that could arise in the other, more specific obligations. The focus on conduct ("use of force") rather than specific technology future proofs the treaty's obligations because it is impossible to envision all technology that could prove problematic. The reference to use of force also allows for application to both situations of armed conflict and law enforcement operations.

Prohibitions

The second category of obligations is prohibitions on weapons systems that select and engage targets and by their nature—rather than by the manner of their use—pose fundamental moral or legal problems. The new treaty should prohibit the development, production, and use of systems that are inherently unacceptable. The clarity of such prohibitions facilitates monitoring, compliance, and enforcement. Their absolute nature increases stigma, which can in turn influence states not party and non-state actors.

The proposed treaty contains two subcategories of prohibitions. First, the prohibitions cover systems that always select and engage targets without meaningful human control. Such systems might operate, for example, through machine learning and thus be too complex for humans to understand and control. Second, the prohibitions could extend to other systems that select and engage targets and are by their nature problematic: specifically, systems that use certain types of data—such as weight, heat, or sound—to represent people, regardless of whether they are combatants. Killing or injuring humans based on such data would undermine human dignity and dehumanize violence. In addition, whether by design or

due to algorithmic bias, they may rely on discriminatory indicators to choose targets.[10]

Positive obligations

The third category of obligations encompasses positive obligations to ensure meaningful human control is maintained over all other systems that select and engage targets. These systems may not be prohibited under the treaty as inherently problematic, but they might have the potential to be *used* without meaningful human control. The positive obligations apply to all systems that select and engage targets based on sensor processing, and they establish requirements to ensure that human control over these systems is meaningful. The components of meaningful human control discussed above can help determine the criteria necessary to ensure systems are used only with such control.

Other elements

The elements outlined above are not the only elements of a new legally binding instrument. While beyond the scope of this chapter, other important elements include:

- A preamble, which would articulate the treaty's purpose;
- Reporting requirements to promote transparency and facilitate monitoring;
- Verification and cooperative compliance measures to enforce the treaty's provisions;
- A framework for regular meetings of states parties to review the status and operation of the treaty, identify implementation gaps, and set goals for the future;

10 For further discussion of the second subcategory of prohibitions, see Article 36, "Targeting People: Key Issues in the Regulation of Autonomous Weapons Systems," November 2019, <http://www.article36.org/wp-content/uploads/2019/11/targeting-people.pdf> (accessed May 21, 2020).

- Requirements to adopt national implementation measures; and
- The threshold for entry into force.[11]

Conclusion

After six years of CCW discussions, states should actively consider the elements of a new treaty and pursue negotiations to realize them. In theory, negotiations could lead to a new CCW protocol, but certain states have taken advantage of the CCW's consensus rules to block progress. Therefore, it is time to consider an alternative forum. States could start an independent process of the kind used to create the Mine Ban Treaty or the Convention on Cluster Munitions, or they could adopt a treaty under the auspices of the UN General Assembly as was done for the Arms Trade Treaty and the Treaty on the Prohibition on Nuclear Weapons. Ultimately, states should pursue the most efficient path to the most effective treaty that preempts the dangers posed by fully autonomous weapons.

11 Campaign to Stop Killer Robots, "Key Elements of a Treaty on Fully Autonomous Weapons," p. 9.

Statement of the Director of the Department of Disarmament, Arms Control and Non-Proliferation of the Ministry for Europe, Integration and Foreign Affairs of Austria

Ambassador Thomas Hajnoczi
Department of Disarmament, Arms Control and Non-Proliferation – Ministry for Europe, Integration and Foreign Affairs (Austria)

Autonomous weapons systems raise unique issues and challenges for IHL compliance from a legal and ethical perspective.

The underlying basis is the reaffirmation contained in the "11Guiding Principles" elaborated by the GGE LAWS that International Law (IL), and IHL in particular, applies to LAWS and that the choice of means of warfare is not unlimited. The human element is critical to IL and IHL compliance. Now the key question is to determine the type and degree of human control necessary to ensure compliance with IL, IHL, the core principles of IHL, and customary IL such as the dictates of public conscience. Legal

obligations, responsibility, and accountability can by definition not be outsourced to machines. International legal norms are based on humans and directed towards humans. States and humans are subjects of law, not machines.

The assessment of compliance with the existing standards and rules under IHL has to be taken contextually in the light of concrete circumstances. Circumstances in the battlefield are shifting, and human control of a weapon is a prerequisite.

There are at least two dimensions to IL and IHL in particular with regard to compliance: **the legality of a weapon *per se*** and the **question of a lawful use of a certain weapon**.

First, **the legality of a weapon *per se***. Means and methods on war are not unlimited. During the development of new technologies, states must ensure that any potential weapon would *per se* be capable to respect basic principles such as distinction, proportionality, and precaution in attack. If a weapon is by its mere design not compatible with IL, it must not be developed. In my view, weapon systems with autonomy in critical functions are a case in point.

IL recognizes the concept of weapons that are indiscriminate by nature due to their unacceptable humanitarian harm. Whether a weapon is potentially lethal or not is not an established criterion under IL. Where would be the added value in introducing such a new category at this point? To be clear, a weapon that delivers lethal effects might very well be used in compliance with IHL.

Second. As we are exploring the limits of the acceptable, the second dimension, the **question of a possible lawful use of a certain weapon system deserves particular attention: What are the key challenges that autonomous weapons systems without meaningful human control over critical functions would pose to IHL**? IHL compliance is highly context-dependent, which is particularly sensitive when it comes to emerging technologies with autonomy

in critical functions. Any use of new weapons needs to comply, *inter alia*, with the **three fundamental IHL principles, namely the principle of proportionality, distinction, and precaution in attack**.

Proportionality requires distinctively human judgement. The assessment must be based on information reasonably available not only at the time of the planning of the attack, but needs to remain valid throughout the weapon's use. The principle of proportionality requires, therefore, an immediate temporal link between the assessment and the factual deployment and use of the weapon. A correct evaluation under the proportionality principle can be a particularly challenging task, for example in populated areas where the situation changes rapidly. Under these circumstances, it would be impossible to weigh anticipated military advantages against the expected collateral harm well in advance. Whether an attack complies with this principle, it is necessary to assess it on a case-by-case basis considering the specific context, the totality of circumstances, as well as the temporal proximity to the attack.

The **principle of distinction** requires distinguishing between combatants and civilians. While it is difficult to assess future technological progress in this regard, substantial concerns exist about data accuracy, bias, and availability of data in conflict situations. It is important to reiterate that, from a legal and ethical perspective, it is more than problematic to leave the selection of targets and decision to attack to a machine; therefore, we cannot envisage how such a system would be compatible with IL. Under the principle of distinction, the respect for the adequate assessment of a person *hors de combat* is equally problematic and requires human judgement.

The **principle of precaution** requires an attack's cancellation or suspension if it becomes apparent that the objective is not a military one, is subject to special protection, or can violate the rule

of proportionality. Besides that, it would be possible for humans to override the system. They all constitute a challenge to LAWS.

In the context of LAWS, ethical considerations are of particular importance. The appropriate legal framework is provided for, *inter alia*, by the dictates of public conscience and the principles of humanity, as referred to in the Geneva Conventions but also in the CCW preamble. IHL is grounded on the basic values of humanity shared by all civilizations. The Martens clause demands the application of "the principle of humanity" in armed conflict.

Ensuring meaningful human control requires a multidimensional approach, which also relates to the **level of predictability and reliability required to ensure human control and the necessary required human legal and situational judgement.** This brings us to the issue of the unpredictability of machine learning algorithms. I agree with the International Committee of the Red Cross's view that "setting boundaries—or operational constraints—in the operation of an autonomous robotic system—for example, on the task, the time-frame of operation, the scope of movement over an area, and the operating environment—can contribute to increasing predictability." Predictability and reliability are crucial for IHL compliance as both contribute to estimating the expected effects and results of a particular weapon's use.

These substantial ethical and legal challenges and concerns bring us to the conclusion that **LAWS without meaningful human control over critical functions would be fundamentally incompatible with IL**.

Lastly, I wish to address the issue of national weapon reviews, also referred to as Article 36 weapon reviews. The objective of these reviews is explicitly mentioned in the guiding principles.[12] There seems

12 See possible Guiding Principle (d): In accordance with states' obligations under international law, in the study, development, acquisition, or adoption of a new weapon, means or method of warfare,

to be a convergence of views on the importance and the merits, but also limitations of Article 36. Weapon reviews constitute a critical national implementation mechanism to establish the legality of a weapon, means or method of warfare. However, Article 36 itself does not give a clear legal standard, it merely assesses whether—from a national perspective—a certain weapon development would be permitted under international law.

Due to military secrecy and development, which is usually seeking a competitive advantage, concrete results and national internal reasoning of a specific Article 36 review are usually not shared with the broader international community. This is closely linked to the **challenge of how states interpret existing norms (including IL, IHL, and the dictates of public conscience)**. If there is no explicit international special norm, states usually differ in their assessment on whether a weapon system is compatible with IL. In the past, in such cases where states felt the need to further clarify international law, more specific regulations were adopted. Under the CCW, Protocol IV is a case in point, where states, given the potential gravity of such weapons being developed, recognized that blinding laser weapons should be prohibited pre-emptively. In the context of LAWS, there is an innate need to internationally clarify the minimum human control acceptable in an autonomous weapon system. A specific international legal norm is thus needed.

Therefore, Austria, Brazil, and Chile have together submitted a proposal to negotiate a new protocol in the framework of the CCW. An increasing number of states concur with the concept of human control as the yardstick and demand a legal regulation. For all who prefer to keep the issue in the CCW, it would be a logical consequence to agree on a mandate for negotiations of an additional

determination must be made whether its employment would, in some or all circumstances, be prohibited by international law.

protocol in this framework. In view of fast technological progress, the work of the GGE-LAWS has to move beyond mere discussions and non-binding guiding principles.

To sum it up, the development and use of an autonomous weapon system have to be assessed on the basis of IL, principles of IHL, in particular the principles of distinction, proportionality, and precaution in attack, and finally in a broader, overarching perspective in its relation to public conscience and the principle of humanity. The necessity to retain meaningful human control is derived from this legal analysis. As Austria's Minister for Foreign Affairs Schallenberg has stated: "We have to regulate LAWS, before they appear on the battle fields."

Panel 3:
Strategic and Military Dimensions of Autonomous Weapons – Disruptive Technology as a Game Changer

Moderator:
Antonio Jorge Ramalho

University of Brasilia

Good afternoon everyone.

I think I should switch to Portuguese, so that we will make it more balanced in terms of the participation of everyone.

I would then like to thank ambassador Candeas once again for organizing this very timely event.

I would also like to thank the Naval War College and the Brazilian Navy for hosting the event and for helping to conceive it, to promote it at such a difficult time.

At a time of important changes in the international scenario, as in any time of major transformations, this brings us many uncertainties, many risks, but also opportunities.

In addition, here you have the opportunity to better understand the positions and thereby propose a solution to a problem that was created by man.

The only clear aspect for all of us is that this is a revolutionary change; this is a revolutionary technology, which will produce transformations very quickly, even faster than any other revolutionary technologies that have transformed our lives, as was the case with the steam machine, as was the case with electricity, as was the case with information technology itself.

We know some things, we know that in these times of uncertainty and great instability, multipolarity is strengthened while the multilateral system is weakening.

We know that the very way of waging war has been questioned. Some say that we are living a fourth generation of war, when certain practices of the past, such as the use of deception, such as the use of complex psychological operations, when the use of terrorist acts were common in war and have been banned since the twentieth century in this civilizing process.

We do not know if this technology will transform war, in its essence, or if it will remain valid.

But fortunately we have here a group of panelists, capable, intelligent, insightful, who will help us understand the meaning of this new technology in contemporary warfare.

The panel deals, as you know, with the strategic and military dimension of autonomous weapons, whether disruptive technology is a game changer and to what extent.

We will follow the order in the program with the Associate General Counsel of the United States Department of Defense, Karl Chang;

Followed by Dr. Roberto Gallo, President of the Brazilian Association of Security and Defense Industries;

Chen Yongcan, Deputy Consul General (China);

After him, Dr. Moa Peldán Carlsson of SIPRI, an institute known to all of us;

And we will conclude with the presentation of Vice-Admiral Alfredo Muradas, Director of Weapons Systems of the Brazilian Navy.

You therefore have before you a very complex issue, some known issues, other issues on which world leaders have also had positioned themselves from an ethical point of view, the ethical issues that were raised here in the panels that preceded us. And that were also raised, for example, with the invention of the submarine, some considered it unethical, an attack coming from someone whom we could not see. The same thing arose with the use of aerial power. Similar issues were also employed when the systems outside the Earth, positioned outside the planet Earth, began to be used to support combat. There too, on all these occasions there were ethical issues, and leaders stated their positions regarding them.

In your opinion, then, let us see if our leaders will also be able to position themselves constructively in relation to this new technological challenge, which, if it is disruptive, brings us some problems already known to humanity.

That said, I immediately give the floor to Councillor Karl Chang.

Implications of Strategic and Military Dimensions of Emerging Technologies in the Area of LAWS for the Work of the GGE Established by the CCW

Karl Chang
U.S. Department of Defense

Introduction and overview

Thank you to the Government of Brazil, Ambassador Candeas, and other Brazilian colleagues for the hospitality and leadership on this issue. Thanks also to the distinguished panelists for their presentations and to everyone for their contributions to these discussions.[1]

1 References attributing remarks to other panelists or participants in the seminar have been removed in accordance with the Chatham House rule. The appearance of external hyperlinks does not constitute endorsement by the DoD of the linked websites, or the information, products, or services contained therein. The DoD does not exercise any editorial, security, or other control over the information you may find at these locations.

I will begin with a few observations on aspects of the strategic and military dimensions of emerging technologies in the area of LAWS. These include:

a. uncertainty about the course of technological development;
b. strategic significance, including "game-changing" disruption; and
c. increased speed, accuracy, and precision in decision-making and the use of force in combat operations.

Second, I want to discuss why these aspects make it difficult to apply to emerging technology in the area of LAWS the disarmament or arms control approaches that have been applied to certain other types of weapons in the past.

Third, I would like to explain how the strategic and military implications indicate that the GGE should focus its work on four related areas:

a. More specific articulations of the requirements of IHL in using weapons with autonomous functions or features;
b. Good practices on human-machine interaction to avoid accidents and to ensure that force is used in accordance with the intention of commanders and the operators of weapon systems;
c. Review processes, such as processes for the legal review of new weapons; and
d. Risk assessments and mitigation measures.

Observations regarding the strategic and military dimensions

A broader context of technological disruption and revolution?

I would like to commend the organizers for the title of this panel, which draws our attention to the fact that **emerging technology**

can have a disruptive, "game-changing" effect that renders obsolete previous ways of doing business, while enabling wholly new capabilities.

There are many examples that we have all encountered in daily life. Video cassette tapes were replaced by DVDs, which are being replaced by streaming video services. This particular example shows that new technologies can improve on the prior generation of technology. There is better picture quality, and it does not take up space on your shelves. It can be less expensive. But new technologies also enable entirely new capabilities. When you use a streaming video service, you are telling that service what you are watching, and it suggests similar things that you might enjoy watching based on your past history and preferences. It can be a little disconcerting at first, but it can be convenient and provide a new capability that was not possible before.[2]

Another observation regarding the strategic and military dimensions that I believe informs the GGE's work is that **there is considerable uncertainty about the future of technological progress**. It often may be difficult to predict how technology will develop or what will be possible. It also may be quite difficult to predict how people will use new technologies. If you consider science fiction depictions from the past, you will see many things anticipated to be commonplace that we do not have today—flying cars, for example.[3] Yet there are also important, transformative technologies

2 For a discussion of potential challenges posed by these sorts of automated viewing recommendations, see Kevin Roose, ***The Making of a Youtube Radical***, The New York Times, June 8, 2019, available at: <https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html>.

3 Assistant Secretary of State Ford makes this observation: "Most of us catch the 'Skynet' references, and some of us who are old enough will remember the rogue computer HAL from Stanley Kubrick's masterful film ***2001: A Space Odyssey***, but it is also true that pop culture predictions of the future have a notably poor track record. Having discovered that we don't actually now live in George Jetson's world—or the world of chauffeured craft swimming through the air to and from the Paris Opera depicted in that marvelous illustration by the 19th-Century French futurist Albert Robida—we should have more intellectual humility than to think we can understand all that much about

and applications that were largely unanticipated—the Internet and social media platforms, for example.

Despite considerable technological uncertainty, it seems very plausible to thoughtful commentators that **technological developments in artificial intelligence and other autonomy-related technologies will rival or exceed human performance at many activities and could lead to widespread changes on the scale of the industrial revolution**.[4] Although states are pursuing military applications of artificial intelligence and other autonomy-related technologies, these technologies are being developed in the commercial sector and are readily available and useful for a variety of non-military purposes.

This type of **change isn't likely to be limited to one state, but could be broadly transformative**. Almost every country in the world has computers and software.

What are the military advantages from AI and Autonomy-Related Technologies?

Military advantages from artificial intelligence and autonomy-related technologies include the following.

First, there can be improved accuracy of decision-making and information-processing. For example, military forces might have surveillance footage from an observation drone, but there might not be enough intelligence analysts to watch all of the many hours of video that have been collected. Just like you can use a search

our technological future." Christopher A. Ford, U.S. Assistant Secretary of State for International Security and Nonproliferation, Arms Control and International Security Papers: AI, Human-Machine Interaction, and Autonomous Weapons: Thinking Carefully About Taking "Killer Robots" Seriously, April 20, 2020, available at: <https://www.state.gov/wp-content/uploads/2020/06/T-Paper-Series-2-LAWS-FINAL-508.pdf>.

4 *See, e.g.*, Henry Kissinger, Eric Schmidt, & Daniel Huttenlocher, ***The Metamorphosis***, The Atlantic, August 2019, available at: <https://www.theatlantic.com/magazine/archive/2019/08/henry-kissinger-the-metamorphosis-ai/592771/>.

engine on the internet to search images, intelligence analysts may want to search video footage to identify where insurgents may have dug holes in the road to plant improvised explosive devices.

A second military advantage is reducing the cognitive load from rote tasks in the operation of machines and allowing more focus on advanced decision-making. Greater autonomy can remove the need for constant input from human operators, which can allow for higher-level control or supervision of multiple unmanned assets simultaneously, and can increase effectiveness by reducing the operator's cognitive load, allowing operators to make command decisions and perform other high-level tasks.[5]

A third military advantage is improved precision and speed in using force in combat operations. For example, consider the C-RAM, the Counter-Rocket Artillery and Mortar system, which the United States has presented on in previous GGEs.[6] This weapon system is a cannon that can shoot down incoming mortars and rockets. Computers, software, and sensors allow the control of a weapon system that is more precise and faster than the manual control of the weapon system by a human gunner.

Observations regarding the military applications of AI and Autonomy-Related Technologies

I would draw four more observations about these military advantages:

First, **these technologies have been used by militaries in some form for many years**. For example, homing missiles with automated target recognition and acquisition systems, and missile

5 U.S. Department of Defense, Office of the Assistant Secretary of Defense for Acquisition and Office of the Assistant Secretary of Defense for Research and Engineering, ***Unmanned Systems Integrated Roadmap FY 2017-2042***, p.20 (footnotes omitted).

6 This presentation is available at: <https://geneva.usmission.gov/2018/04/13/ccw-gge-u-s-slide-presentation-counter-rocket-artillery-and-mortar-system-c-ram/>.

defense systems like the AEGIS system,[7] have been fielded and used for decades.

Second, **whether weapons are characterized as "autonomous" can depend on how the system is used, rather than the intrinsic characteristics of the weapon system**. For example, consider a missile with automated target recognition capabilities that can select and engage enemy tanks. In one scenario an operator identifies a specific target and fires the missile at this target. Under the definitions applied by the U.S. military, this is a semi-autonomous weapon system.That same weapon system and capability could, however, be classified as an autonomous system if it is used in a different way. If the operator does not identify a specific tank, but instead fires the weapon to loiter in an area and autonomously select and engage tanks, the weapon is classified as an autonomous weapon in U.S. military practice. The point I am trying to illustrate is that the weapon system's technical characteristics are the same, but how it is to be used changes whether it is classified as autonomous or semi-autonomous. The question I would pose to those who have concerns about autonomous weapons is whether these concerns are really about a type of weapon system or whether they are about how weapons systems are used.

Third, the **reliance on autonomous functions to aid in decision-making might not be intrinsically part of the weapon system**. For example, as the United States has discussed in the GGE, counter-battery radar systems can be used to identify the location from which incoming fire originates.[8] These systems then can be used

7 U.S. Navy Fact File: AEGIS Weapon System, Jan. 10, 2019, available at: <https://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=200&ct=2>.

8 Shawn Steene, Member of the U.S. Delegation to the Convention on Certain Conventional Weapons Group of Governmental Experts on emerging technologies in the area of lethal autonomous weapons systems (LAWS), U.S. Practice in the Assessment of Weapons Systems, Geneva, March 27, 2019, available at: <https://geneva.usmission.gov/2019/03/28/convention-on-ccw-u-s-practice-in-the-assessment-of-weapons-systems/>.

to direct counter-battery fire. This is not a weapon system as such, but is an application of technology that informs human-decision making in combat operations.

Consider this analogy. You can use automation to help you steer a car. You also can use artificial intelligence to tell you where to drive—with a mapping application that tells you how to travel with the least amount of traffic. The mapping application is on your phone; it is not necessarily in your car. Similarly, the emerging technologies that might be very relevant to the use of the weapon system might not be part of the weapon system.

My fourth point is that military advantages from **these technologies can enhance implementation of IHL in military operations, such as reducing the risk of civilian casualties**. Such humanitarian benefits may include, for example, increasing awareness of civilians and civilian objects on the battlefield, and reducing the need for immediate fire in self-defense.

In many instances, civilian casualties are caused because commanders and operators were not aware of the presence of civilians and civilian objects. Use of AI can help improve situational awareness and the detection of civilians and civilian objects. People have raised concerns about the use of AI or autonomy to identify individuals, but this technical capability actually could help reduce civilian harm.

Another situation in which civilians are at increased risk is when military forces are in contact with the enemy and need to respond to enemy fire in self-defense. In those operational situations, the imperative to take immediate action to counter a threat from the enemy reduces the time available to take precautions to reduce the risk of civilian casualties.

Existing practice, however, suggests that emerging technologies may offer ways to reduce civilian casualties in this situation. First,

the use of robotic and autonomous systems can reduce the need for immediate self-defense fire by reducing the exposure of human beings to hostile fire. For example, remotely piloted aircraft or ground robots have been used to scout ahead of forces conducting patrols in environments where they might be surprised by enemy ambushes or roadside bombs. Robotic and autonomous systems can provide a greater standoff distance from enemy formations, allowing forces to exercise tactical patience to reduce the risk of civilian casualties.

Second, technologies to automatically identify the direction and location of incoming fire can reduce the risk of misidentifying the location or source of threats.

Third, the use of defensive autonomous weapons used to counter incoming rockets, mortars, and artillery can provide additional time to develop a considered response to an enemy threat as opposed to responding immediately with counter-battery fire.

The United States discussed these and other humanitarian benefits in a working paper that we submitted to the GGE in 2018.[9]

How strategic and military dimensions can inform the work of the GGE

Now I would like to discuss how these strategic and military dimensions could inform the work of the GGE.

Novel aspects of LAWS make applying traditional disarmament approaches questionable

First, I would like to pick up on a point that was mentioned at the opening of our seminar about the differences between LAWS and other types of weapons that have been the subject of arms

9 U.S. Working Paper, *Humanitarian Benefits of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, March 28, 2018, CCW/GGE.1/2018/WP.4, available at: <https://www.unog.ch/80256EDD006B8954/(httpAssets)/7C177AE5BC10B588C125825F004B06BE/$file/CCW_GGE.1_2018_WP.4.pdf>.

control. In light of those differences, it is not clear that traditional disarmament approaches can be applied to LAWS.

Moreover, rather than being analogous to weapons with little military importance (*e.g.*, weapons that injure by fragments non-detectable by x-ray prohibited by CCW Protocol I), these technologies may have potential "game-changing effects" and are currently employed in many military systems, such as combat aircraft, warships, and missiles.

The commercial sector's development of these technologies and their many non-military purposes suggest that the underlying components (*e.g.*, computers, software, and sensors) that distinguish autonomous weapons are not easily subject to traditional arms control restrictions, either.

The potential for emerging technologies to reduce the risk of civilian casualties also counsels against simply seeking to ban these technologies or their military applications. These are technologies that can be used to create more discriminate effects.

Four promising areas for the GGE's work

Although traditional disarmament approaches may not be successful, the GGE can usefully elaborate and develop its work in four areas, at least:

a. a better understanding of IHL requirements;

b. good practices for human-machine interaction;

c. review processes; and

d. risk assessments.

First, uncertainty about the course of technological progress does not affect IHL, which, as guiding principle (a), adopted by the GGE in 2019, recognizes, continues to apply fully to all weapons systems, including the potential development and use of LAWS.

The 2019 GGE focused on IHL, adding an agenda topic and nine consensus, substantive conclusion paragraphs on this issue. More work on IHL is possible to clarify IHL requirements. In its 2019 working paper, the United States discusses what IHL requires in three use scenarios.[10] The GGE should continue to develop better common understandings of what IHL requires when using emerging technologies in the area of LAWS. Focusing on more specific use scenarios is perhaps one way to develop a clearer and more granular discussion.

Second, this work on IHL should be informed by discussions on good practices for human-machine interaction. As guiding principle (c), adopted by the GGE in 2019, recognizes, human-machine interaction "may take various forms and be implemented at various stages of the life cycle of a weapon,"[11] and a range of factors should be considered in determining the quality and extent of human-machine interaction.

The GGE can usefully elaborate upon these factors and potential good practices. The United States would welcome the opportunity to submit our practice in this area and to learn from the practice of other states.

Third, review processes allow decisions to be made on a case-by-case basis in the particular circumstances and thus can help address the technological uncertainty surrounding emerging technologies in the area of LAWS. The GGE also has emphasized guiding principle (e), which restates the obligation to conduct legal reviews of weapons

10 U.S. Working Paper, ***Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems***, March 28, 2019, CCW/GGE.1/2019/WP.5, available at: <https://www.unog.ch/80256EDD006B8954/(httpAssets)/B2A09D0D6083CB7CC125841E0035529D/$file/CCW_GGE.1_2019_WP.5.pdf>.

11 Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, U.N. Doc. CCW/GGE.1/2019/3, ¶16, Sept. 25, 2019, available at: <https://undocs.org/en/CCW/GGE.1/2019/3>.

found in Article 36 of the 1977 Additional Protocol I to the 1949 Geneva Conventions.[12] Although the United States is not a party to the 1977 Additional Protocol I, the U.S. military engages in a practice of reviewing weapons before they are acquired or procured to ensure their consistency with applicable international law.[13] We think it would be productive for the GGE to seek to compile good practices in the legal review of weapons systems based on emerging technologies in the area of LAWS.

Fourth, risk assessments and mitigation measures provide another way to help address uncertainty and to balance competing risks and benefits. Japan's practice in this regard was discussed earlier today, including Japan's practice regarding safety requirements for personal care robots as well as Japan's AI utilization guidelines. The idea is to seek the benefits of emerging technologies but also to take deliberate steps to minimize risks.

The GGE's guiding principle (g), which also was highlighted for us in the first panel, provides that "Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems,"[14] and paragraphs 23(a) and (b) of the GGE's 2019 report discuss types of risks to be considered and mitigation measures.[15] Further work on risk assessment processes and mitigation measures could allow

12 Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, U.N. Doc. CCW/GGE.1/2019/3, Annex IV ¶(e), Sept. 25, 2019, available at: <https://undocs.org/en/CCW/GGE.1/2019/3>.

13 For a discussion of U.S. Department of Defense practice in weapons reviews, see Department of Defense Response to Stockholm International Peace Research Institute (SIPRI) "Questionnaire on Article 36 Review Process", Sept. 1, 2017, available at: <https://ogc.osd.mil/LoW/practice/DoDDocuments/sipri_questionnaire_on_article_36_review_process_usa_response_final.pdf>.

14 Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, U.N. Doc. CCW/GGE.1/2019/3, ¶(g), Sept. 25, 2019, available at: <https://undocs.org/en/CCW/GGE.1/2019/3>.

15 Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, U.N. Doc. CCW/GGE.1/2019/3, ¶23, Sept. 25, 2019, available at: <https://undocs.org/en/CCW/GGE.1/2019/3>.

the GGE to provide practical and implementable recommendations for states to address concerns.

These areas that I have mentioned, articulating IHL requirements with more specificity, good practices on human-machine interaction, review processes, and risk assessments and mitigation measures, are very much inter-related. This morning a very thoughtful question was asked about the limits of weapons reviews. And I want to address that question in the spirit of building bridges. Weapons reviews are not alone the solution, but when combined with other measures as part of a framework, the utility of weapons reviews becomes more apparent. These reviews provide states the opportunity to consider issues raised in terms of implementation of IHL at an early stage of the work on a given capability. For example, in legal review processes, we assess risks, such as the types of risks that the GGE has identified. We consider how to comply with IHL and whether there are good practices in human-machine interaction that can help ensure compliance with IHL and can mitigate risks. I would encourage the GGE to consider how the individual aspects of its work, like weapons reviews, can work together as a part of a framework to address emerging technologies in the area of LAWS.

Concluding thoughts

In conclusion, I would just like to emphasize a few points.

This issue, from the perspective of the United States, is an important one. States are going to use emerging technologies in the area of LAWS in military operations. How do states use these technologies responsibly, ethically, and in accordance with IHL?

The GGE is an incredible opportunity to have all States that are willing to participate, as well as civil society organizations, discussing this issue and, borrowing the language of the GGE's

mandate, clarifying, considering, and developing aspects of the normative and operational framework.[16]

The United States wants the GGE to be successful. The GGE has made tremendous progress in the 11 guiding principles, but also in the substantive conclusions that have been adopted already by consensus. The U.S. delegation is ready to work constructively to continue that progress over the next two years under the leadership of Ambassador Karklins and others.

16 Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Final Report, ¶31, U.N. Doc. CCW/MSP/2019/9, Dec. 13, 2019, available at: <https://undocs.org/CCW/MSP/2019/9>.

Kill Switch, Switch to Kill: Reflections on Autonomous Weapons Systems and their Impacts on Defense

Roberto Gallo
President at ABIMDE and CEO at Kryptus[1]

Thiago Carneiro
Head of DIPROD/MRE[2]

As each day goes by, military capacities are more dependent on platforms of sophisticated and automatic weapons. Such undeniable technical advances may bring overwhelming advantages to the battlefield, dangerously increasing the gap between countries that develop technologies and those that merely utilize them. However, this phenomenon also implies big strategic risks, typically misunderstood by the makers of military doctrines and by field operators, particularly in countries that only buy those technologies and equipment.

1 Brazilian Defense and Security Industries Association; Kriptus – Solution Povider for Information Security, <gallo@kryptus.com>.

2 Division for Products on Defense – Brazilian Ministry of Foreign Affairs, <thiago.carneiro@itamaraty.gov.br>.

More and more, these changes present themselves as challenges to every armed force, and, equally so, to the defense industry. In order to better comprehend those risks, it is necessary to understand how the automation of combat platforms and its systems have exponentially increased over the past 40 years.

Automation in platforms

If we observe the platforms of the eighties, we notice their embarked components and subsystems operated in a mainly isolated way. Airplanes, for instance, had subsystems such as communications links, mission computers, countermeasures, weapons systems, navigation and motor control, and some had at best some sort of shared reading in the cabins.

Although that type of organization—despite assisting with the control of critical areas—still imposed a major workload on the crew, it also kept the platform in hand. One failure of an individual component could frequently be mitigated by the crews themselves, or even during ground maintenance.

Still using airplanes to exemplify the matter, four decades later, we may observe the contrast with the case of the Boeing 737 MAX 8. The level of automation, integration, and independence of the aircraft is such that the architecture of systems is conceived in a way that the airplane may fly safely *despite* the crew. The human operator, who used to be crucial for the aircraft to function, is increasingly growing to be a supporting character in the (limited) platform control, Figure 1.

Figure 1: Airbus tests a first fully automatic takeoff based on computer view. Part of the ATTOL program

Source: Airbus webpage, accessed on Feb. 23, 2020.

In the military world, that level of automation has led to the autonomy of weapons systems (LAWS[3]) which, supported by artificial intelligence, communications and advanced sensors (C4ISR[4]), have the ability to plan missions, identify and acquire targets, perform lethal actions, and withdraw, with no need for human intervention.

This increasing level of complexity of systems and platforms presents a series of challenges to the defense scholar, on multiple levels:

- Strategic (mastering technical knowledge and geopolitical implications);
- Tactical (reformulation of military doctrines);
- Operational (use of technology in combat and denial of its use to enemies); and

3 Lethal autonomous weapons systems.
4 Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance.

- Moral (how far does the operator's responsibility go? And who is the operator to be held responsible?).

In this article we will address, in a very brief manner, each of these aspects, with particular emphasis on the operational question—in which the concepts of "kill switch" and "switch to kill" stand out—and its implications to Brazil.

Strategic aspect

Eminence in war, and therefore victory at an armed conflict has always been associated with ability, quantity, and mastery of combat means. It has long been known that it is not enough to have the best military strategy if you do not have the support of an important technical and logistic chain. Steel forges, powder stock, observation balloons, submarines, long-haul aircraft, and long-range missiles have all been, each at a time, key elements for supremacy.

If, during the whole history of conflict, supremacy has marched alongside technical superiority, only now do we witness a paradigm shift in the man-weapon relationship. Even with all the evolution of war during the last millennia, systems and platforms have always somehow been under the operator's command. We now witness a revolution as quiet as it is scary. The weapons systems no longer rely on a person in order to be effective during combat.

There is a "depersonalization" of the soldier, who today has gone from operator to system manager, and, in the future, will practically be a simple spectator. That goes, of course, for countries that possess such critical technologies. As for the others, there will always be a place for field operators, or even for "cannon fodder". Such changes profoundly alter the strategic prospect. The world is divided between countries that can make war as offensive as it can be at a minimum cost of their citizens' lives, and those that will suffer the consequences of a clash against an adversary who never

goes hungry, thirsty or tired, and can function 24/7 in the most hostile environments.

Tactical aspect

Similarly, the military doctrines present themselves according to the systems and platforms of their time. From the Roman "turtle" formation, able to use their shields and spears for sustained attack of the legions, to Nazi Germany's "Blitzkrieg," to the expedited advance of "*panzer*" divisions and air support from "*stukas*," up to USA's "Shock and Awe" doctrine, war has always been fought with troop risk mitigation and greatest possible damage imposition to the adversary in mind.

Autonomous platforms emerged for a new generational leap to these doctrines. 21st-Century war centers around nets, intensive use of technology, and lower human involvement to those who can afford to do so. "Drone swarms" and nearly automated armies are a new reality that will significantly alter how states will start, win, and lose wars. This is the context in which three concepts must be well understood by the makers of public policy and military doctrines. The essentiality of the national Defense Industrial Base (DIB) principle as a final means of effective protection of the country's technical sovereignty, which is connected to the strategic and tactical aspects, and concepts like "kill switch" and "switch to kill" are related to the operational aspect.

Operational aspect

The use of autonomous weapons systems in combat reveals a great deal of challenges to the modern soldier. Decision-making in combat—a fundamental element to the success of any maneuver—becomes increasingly less dependent on men, significantly changing the way armies will define their strategies in combat. Autonomous weapons systems may "talk" among themselves, implementing

the true concept of "netcentric warfare," by means of encrypted communications. The ability to process and fuse data, together with the dissemination of information to the attack units, is the difference between victory and defeat. More precise and refined algorithms, modeled by supercomputers, will not only better aid the generals of the future, but, in some cases, replace them altogether.

Whoever has the best combination of software and hardware will have the best ability. However, these changes on how to "make war" lead to two conceptual matters that are as fundamental as they are complex: how to re-establish domination of man over machine in extreme situations and how to make (or avoid) the machine turn on its user. That is where we have the concepts of "kill switch" and "switch to kill," which we will begin discussing now.

Kill Switch

In English, "kill switch" is the term that refers to the command, key, button, or any resource that shuts down or disables a system when so decided. Kill switches *normally* work for the benefit of the system's owner, as a fast stop resource, for protection. Some examples of that are panic buttons in lathes and the remote wipe of an iPhone's data in case it is lost or stolen.

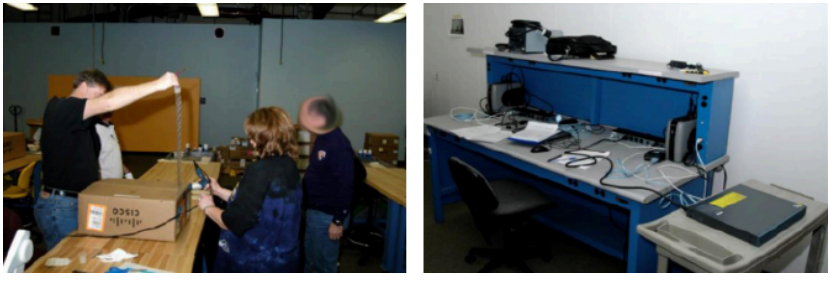
Any given system may have many "kill switches", some activated in person, and others remotely, including through the narrowest of communication bands, after all, one single bit of information is what needs to be transmitted. "Kill switches" may be clearly identified as a "big red button" or insidiously buried in a single electronic component among thousands that may form a complex system—sometimes without the knowledge of the system's manufacturers themselves!

The situation is so serious and the challenge so big that General Keith B. Alexander, former director of the feared National Security Agency—the NSA—reputed the profound "kill switches," inserted

with no knowledge of a technology's owners, as one of the biggest challenges of computer platforms in the future.

Alexander fully comprehended what he was talking about: during the 2010s (after revelations by Edward Snowden, Figure 2), until recently in February 2020 (the case of the company Crypto SG, run by the CIA [Central Intelligence Agency] and BND[5] [Germany's Federal Intelligence Service]), the United States were caught many times practicing what is referred to as "supply chain interdiction."

Figure 2: The NSA systematically changes "hard targets" equipment to compromise their functions in a preparation action called "supply chain intervention attacks"



Source: Edward Snowden leaks

That category of attacks serves as preparatoryaction for purposes of intelligence, creation of "distractions," sabotage, and use denials. While, for intelligence activities, examples of "supply chain interdiction" are recurrent—take the case of Huawei with their 5G technology banned from many countries throughout 2019 and 2020—other finalistic activities are way less reported, but no less real or less impactful.

5 Available at <https://www.washingtonpost.com/graphics/2020/world/national-security/cia-crypto-encryption-machines-espionage/>.

As for the military scope, any minimally equipped nation has formal or informal concerns about *technical denial* and *end-user* agreements that limit acquisitions, information, and also possible employment of certain technologies. The final form of use control, however, is operational impossibility, classically imposed on missiles and combat systems that respectively require launch codes or "software licenses" to operate.

As a classic denial example in our surroundings, we could (and should) mention the one inflicted by the French onthe Argentinians during the Falklands War against the British in the 1980s. At the time, even without launch codes, the Argentinian Air Force was able to deploy EXOCET missiles and surprise the British Navy.

However, that kind of scenario would never repeat itself with today's levels of platform automation—we may suppose, far from fiction, that if the Falklands War took place today, "kill switches" on French fighter-bombers would not even allow the aircrafts to take off, let alone that missiles be launched.

That type of ability today, to remotely deny use through hidden "kill switches" spread across platforms, implies massive strategies, some of them mentioned ahead.

However, there are still some even more serious outcomes when you combine high levels of automation and *supply chain intervention* attacks.

Switch to Kill

Not meaning to sound repetitive, but the obvious should be mentioned: the same control technology that is used to disable a system against the owner's volition may also be used to enable self-destruct routines.

For example, a Trojan Horse in hardware, inserted in the logistics chain of a missile, activated by satellite, may be used not

only to disable the weapon, but to detonate it in the hands of their owners, whenever it is the most damaging. Whether at the bunker or embarked on the platform, the potential damage is enormous.

However, the worst is yet to come: with the advent of LAWS the subject becomes even more complex. Named "killer robots" by human rights activists, the employment of such systems has raised many discussions among the international community, as the technologies have been employed on real operations by the leading nations of the war industry, Figure 3.

Figure 3: Have Raider Demo of Lockheed Martin on LAWS in action with: mission planning and replanning, execution of tasks given by the operator, autonomous ground and formation flying

Problem: The global landscape is rapidly evolving and information in the battlespace is moving faster than ever. In this dynamic environment, it's critical for the warfighter to act on information as soon as it is available. By leveraging the strengths of both systems, manned/unmanned teaming enables rapid action during combat. Successful teaming also has the potential to increase situational awareness and improve combat efficiency and effectiveness for the warfighter.

Requirements: To successfully prove manned/unmanned ground strike teaming capabilities, the unmanned combat air vehicle needed to complete four critical tasks:

- Automatically plan a ground-attack mission
- Execute tasks based on priorities given by the operator
- Dynamically replan the mission to minimize exposure to the threat
- Demonstrate autonomous formation flying, route following and rejoining

Solution: The U.S. Air Force partnered with Lockheed Martin Skunk Works® to tackle this hard problem. The experiment, called Have Raider, was designed to demonstrate the technologies required for an unmanned vehicle to fly as a teammate with a manned vehicle in the battlespace. Using an experimental F-16 as a surrogate unmanned aircraft, the demonstration proved the ability to autonomously plan and execute air-to-ground strike missions, fly in formation with a manned aircraft and react to changing threat environments.

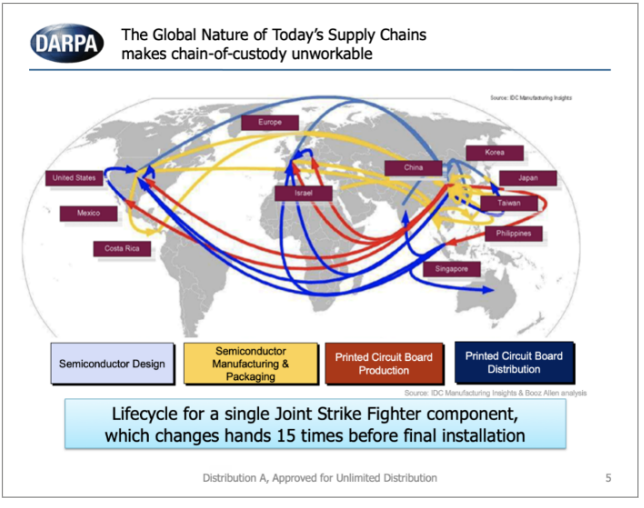
In a military context, however, the broad discussion that is lacking is around the fact that a "kill switch" may be a "switch to kill," a plausible situation in which target designation of a LAWS is subverted, and, instead of attacking foes, it could turn against friends.

As LAWS, by definition, possess minimally autonomous capacities to designate targets and shoot, they also become perfect

targets (or means) to accomplish actions that go *against* the weapons systems owner's interests: (i) supplier/developer nations have great geopolitical and strategic incentives to demand their war industries to include "kill switches" and "switches to kill" in weapons systems supplied to other states, (ii) at the same time, different supplier nations may want to compromise the supply chain of their adversaries.

In fact, the matter is so grave that private statistics demonstrate that about half the suppliers of the American DoD had already faced problems of supply chain intervention in 2015.[6] No wonder the USA has SHIELD (Supply Chain Hardware Integrity for Electronics Defense), the great program led by DARPA (Defense Advanced Research Projects Agency) to mitigate supply chain interventions.

Figure 4: Suplly chain intervention of electronic components mapped by SHIELD/DARPA



Source: Software and Supply Chain Assurance Winter Forum 2018.

6 CHASE Workshop on Secure/Trustworthy Systems and Supply Chain Assurance, University of Connecticut.

Strategic implications

The strategic developments of the aforementioned reality are wide-ranging and deserve long reflection, particularly when it comes to those of second and third order. That being said, the following list should only serve vpurposes of summoning or starting material for studies in Schools of War.

Noting that reservation, we call brief attention to the following:

5. **The essentiality of the national Industrial Defense Base principle** as a vector of technology development and fundamental component in maintaining real sovereignty and the ability to have effective military deterrence. Without the "know how" and "know why" of these technologies, the complex weapons systems are just expensive toys in the taxpayers' eyes and, to some extent, threats to the country's sovereignty;
6. **Higher instability in conventional conflicts**. Because they require less trained military personnel and may be controlled remotely, LAWS can be supplied to a larger number of countries, including in buying options, such as "leasing" or rental (real "mercenary robots"), posing particular risk to civilians;
7. **Catastrophic reduction of strategic employment hypothesis**. No one should expect that a highly automated weapons system (LAWS or not) purchased from a given country could correctly operate against said country's interests, right? However, it is even worse than that: with foreign LAWS, said platform could even fight against those who purchased them (how about that?); and
8. **Acquisitions of defense materials should be widely revised**. While central powers have great programs of supply

> chain assurance to mitigate and control the risks of "kill switches" (for example the American programs Trusted Foundry and DARPA SHIELD), this subject cannot even be comprehended in many other countries. Besides the elementary and necessary solution of strengthening the national industry, it is important to make plans so that, in the strategic aspect, the choices of technology partners favor control and visibility over the subcomponents of military platforms and interoperability among singular forces. We should also pose the question: at the end of the day, who benefits from occasional purchases?

In Brazil, the subject of supply chain protection has gained some traction, even if it is not in a structured manner. On one hand, Bill no. 12,598/2012 collaterally covers the subject. On the other, recent Decree no. 10,222/2020, which establishes the National Strategy of Cybernetics, offers more advance—and it has already reflected on the limitation of Huawei's 5G critical infrastructure in Brazil.

However, facing the problem directly has still not resulted in strategic action, unfortunately, even if the subject has been under discussion for a few years within the technical scope of the Armed Forces and the Brazilian Intelligence Agency.[7] We must establish a policy and a national system of supply chain assurance for national defense, because strategic actions of intervention have already been implemented by the great powers for years.

As for the matter of reorganizing the purchase process, we must initiate an open and clear debate about the current Brazilian template, particularly when we consider the great changes autonomous weapons systems impose on the future. The purchasing of defense material, thought and made only by singular Forces, or with minimal

7 Available at <http://www.abin.gov.br/conteudo/uploads/2018/12/RBI-13_VERSÃO-ELETRÔNICA-Completa-12-12-2018.pdf>.

intervention from other areas, reveals a short-sightedness of action and a culture of operations that is still too compartmentalized, which is unfit for the new setting of conflicts in which the mastery of technology should be a priority.

A new multidisciplinary, inter-ministerial and inter-agency structure is not only desirable, but it also seems to be the only possible way to handle these challenges. The eventual creation of a Special Inter-Ministerial Secretariat (SEIPRODE), which encompasses all the state actors involved with the defense product areas (Ministry of Defense, Ministry of Foreign Affairs, Ministry of the Economy, Ministry of Science, Technology and Innovation, and Brazilian Trade and Investment Promotion Agency) to deal with the subject of defense products from their conception, industrial development, exportation, control, and financing, in constant and fluid contact with the private sector, presents itself as an urgent demand, even if it offers a mid-term resolution.

Moral aspect

In the moral aspect, there are more questions to be asked about the combat operations. Who is in charge of the actions performed by a machine? Who is the operator in charge? How about humanitarian issues, under international law, such as the Geneva Convention, if now the machines decide what, when, and how to destroy? Given the goal of this article, we will not delve into the subject, but we already observe the necessity to reflect on the moral implications—and international law—that the use of autonomous weapons brings.

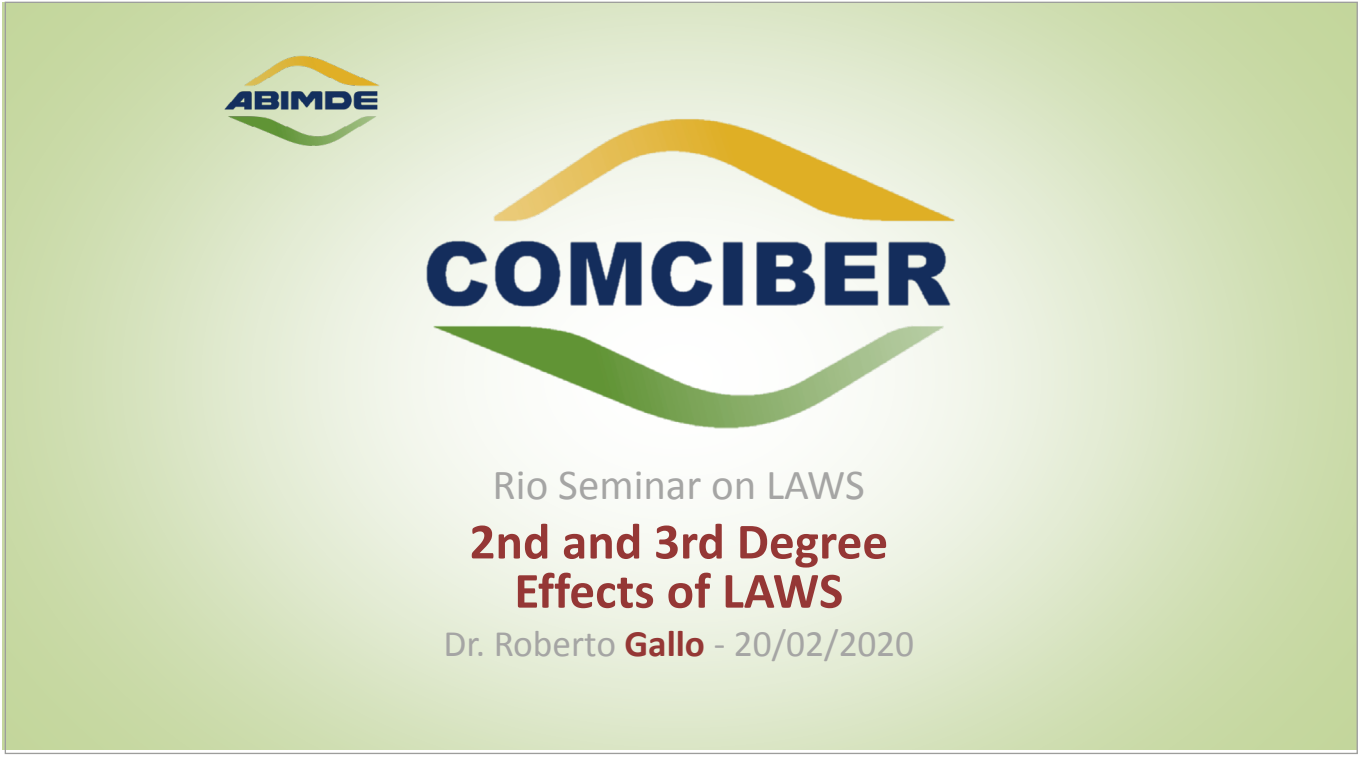
Conclusions

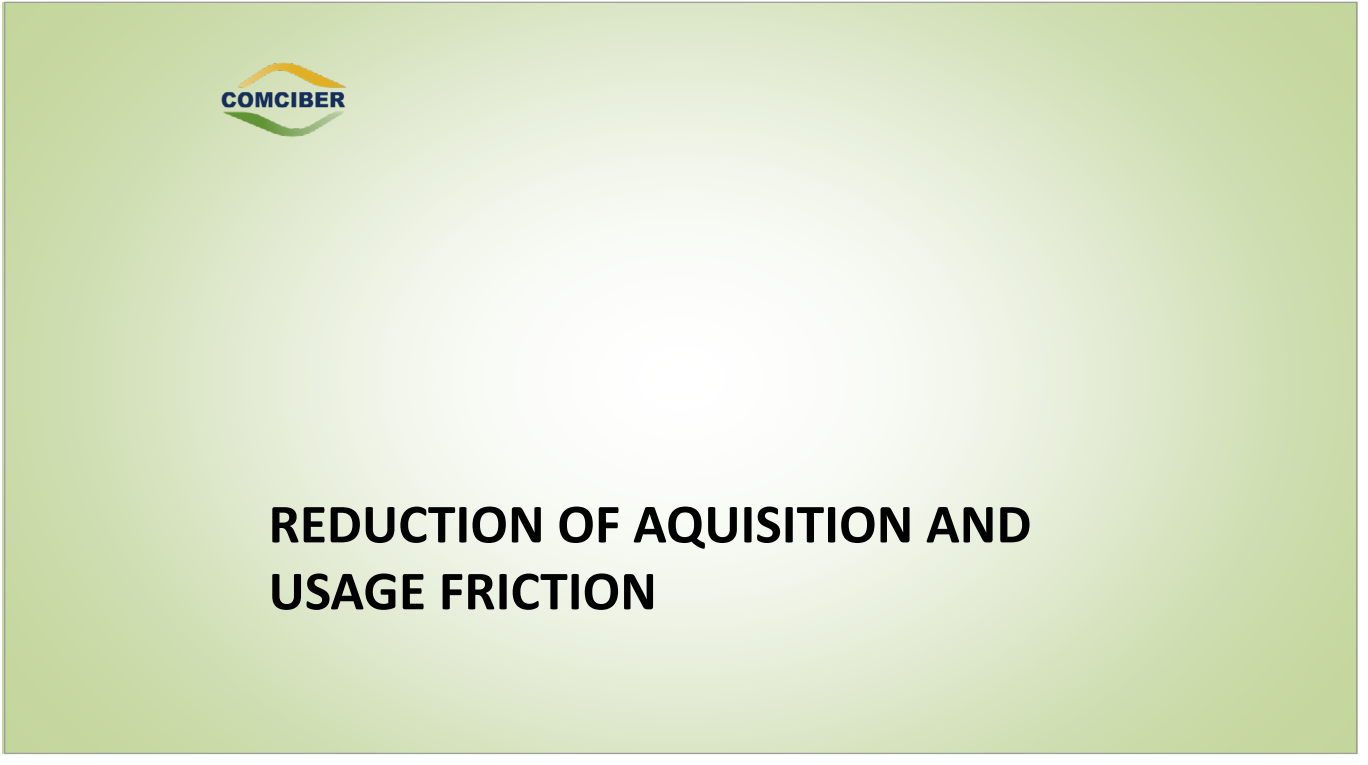
There is a revolution in place regarding weapons systems. Based on technology, such changes imply disruptive effects for national defense and security. In order not to let our national states appear in

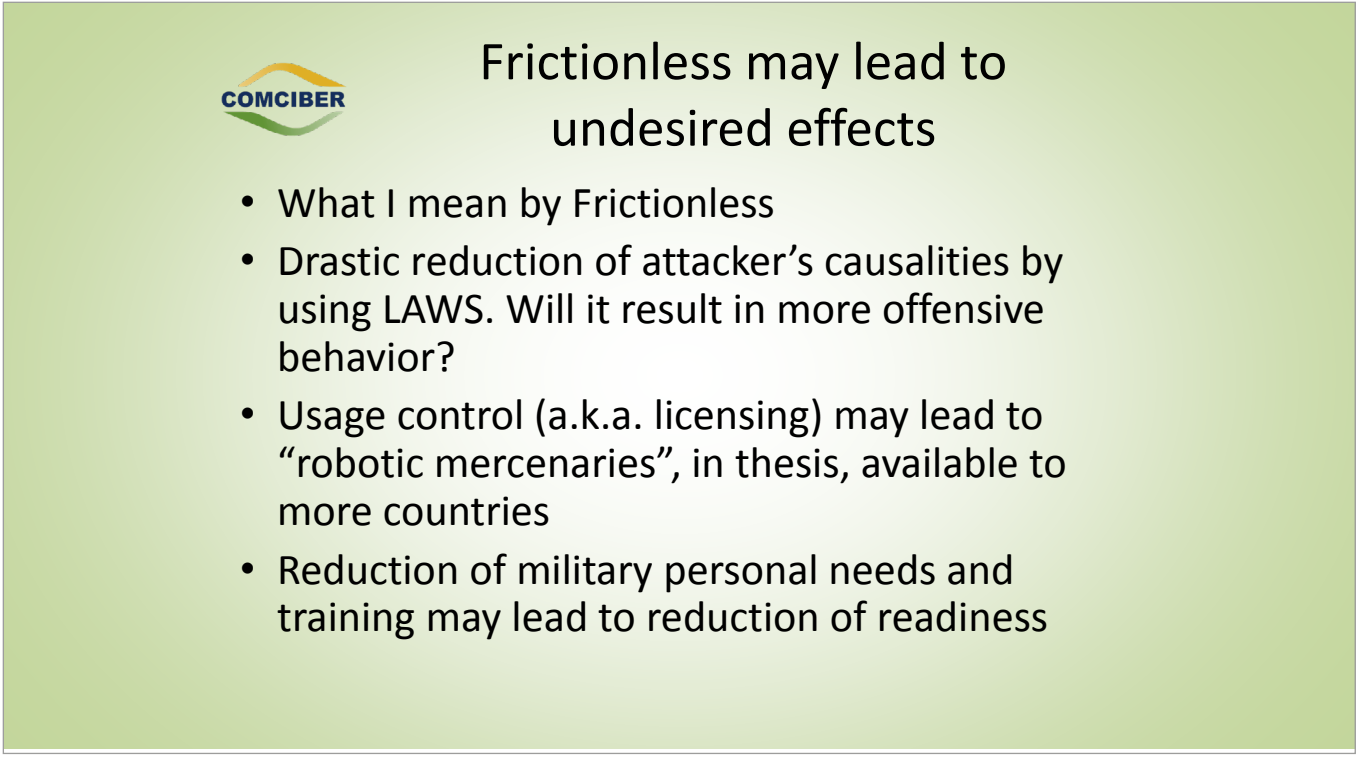
strategically unfavorable positions, it is important to fight technology captivity and recognize the essentiality of their respective DIBs.

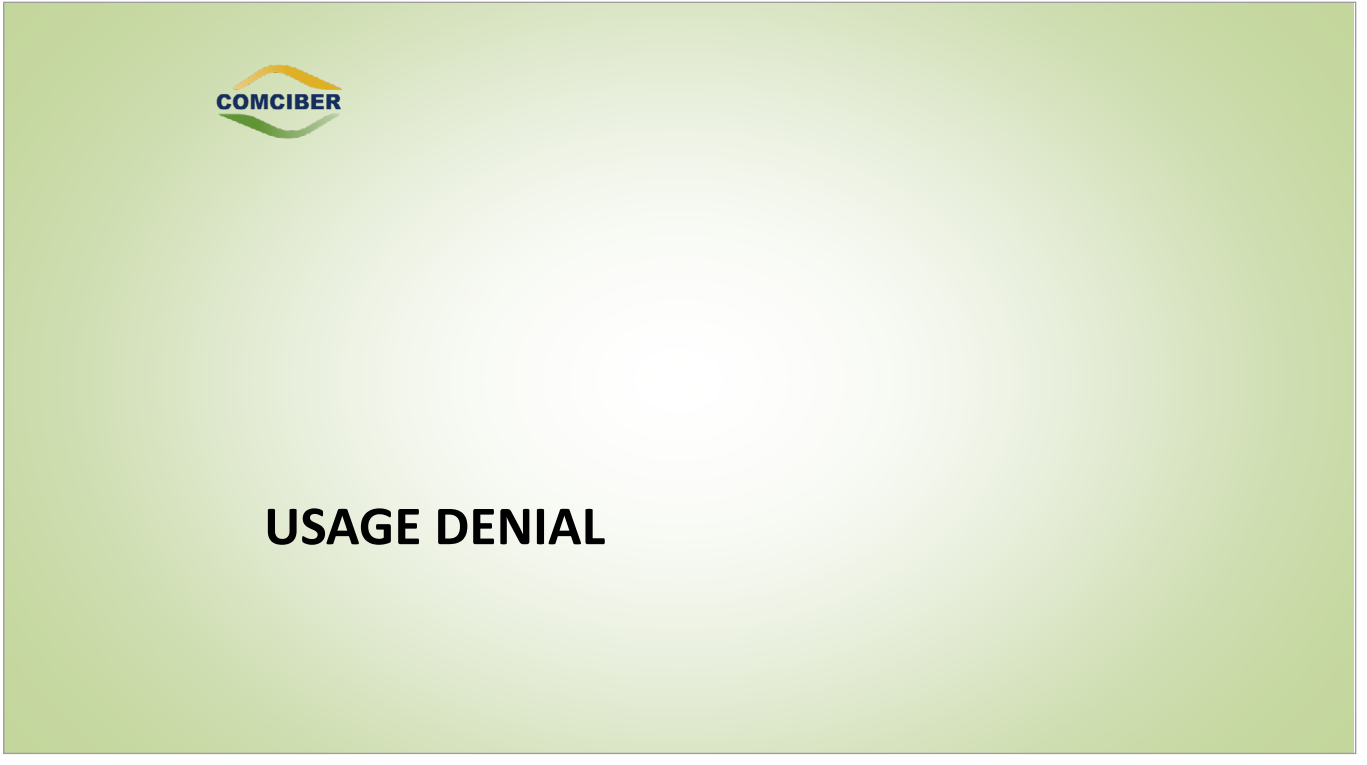
In light of the emergence of LAWS, it is crucial that in Brazil, as well as in other countries of similar stature, we advance in studies of strategic, tactical, operational, and moral implications, which should serve as subsidies for the necessary realignment of national policies and state agencies.

In comparison to other countries, particularly to Turkey, to us it seems that, in the Age of LAWS, Brazil cannot do without a new state structure, one that is multidisciplinary, inter-ministerial, and inter-agency, so that we can keep up with the revolution in place, considering the prominence the country deserves.


ABIMDE
COMCIBER
Rio Seminar on LAWS
2nd and 3rd Degree
Effects of LAWS
Dr. Roberto Gallo - 20/02/2020


COMCIBER
REDUCTION OF AQUISITION AND
USAGE FRICTION


COMCIBER
Frictionless may lead to undesired effects
• What I mean by Frictionless
• Drastic reduction of attacker's causalities by using LAWS. Will it result in more offensive behavior?
• Usage control (a.k.a. licensing) may lead to "robotic mercenaries", in thesis, available to more countries
• Reduction of military personal needs and training may lead to reduction of readiness


COMCIBER
USAGE DENIAL

COMCIBER

Supply Chain Interventions

- Problems in Aerospace Industry
- 173 plane crashes related to supply chain interventions
  - On average 2% two components of an aircraft are frauded
  - Electronics with "backdoor" found in the F-16, C-17, and others!
- 46% of DoD/US suppliers had problems!
  - Discovered: 11% after launch, 32% during production, 28% during prototype
- DARPA – SHIELD Program (with the help of the NSA :-)

COMCIBER

DARPA

OUR WORK

AEO BTO

Microsystems Technology Office

THRUST AREAS PERSONNEL

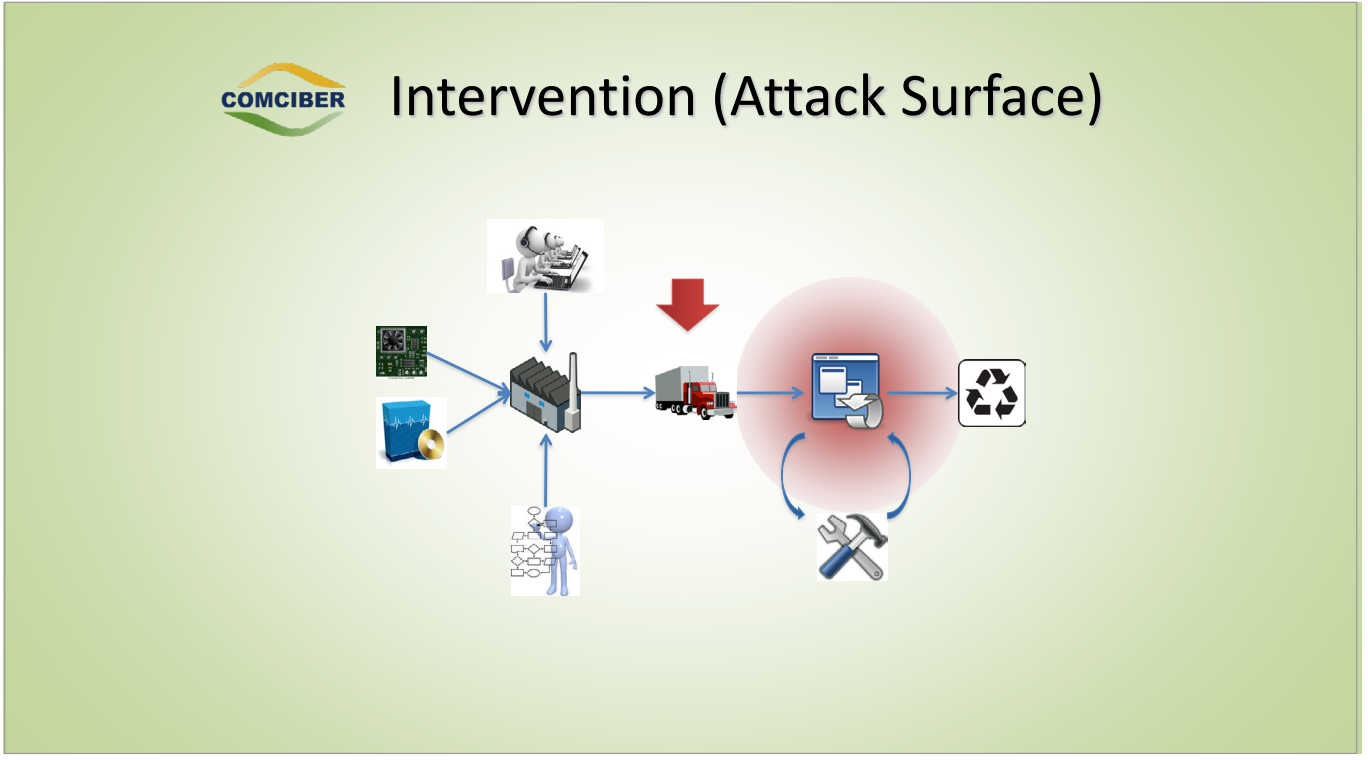
PROGRAM MANAGER

Mr. Kerry Bernstein
kerry.bernstein@darpa.mil

SUPPLY CHAIN HARDWARE INTEGRITY FOR ELECTRONICS DEFENSE (SHIELD)

The security and integrity of DoD electronic systems is challenged by the presence of counterfeit integrated circuits (ICs) in the supply chain. Counterfeiters use a variety of easy and inexpensive techniques to recycle discarded ICs, alter them, and reintroduce them to the supply chain for profit. These parts have questionable reliability and may not function as specified. The failure of a fielded DoD system due to the presence of a counterfeit IC can jeopardize the success of a mission and put lives at risk.

The goal of DARPA's SHIELD program is to eliminate counterfeit integrated circuits from the electronics supply chain by making counterfeiting too complex and time-consuming to be cost effective. SHIELD aims to combine NSA-level encryption, sensors, near-field


COMCIBER
Intervention (Attack Surface)

COMCIBER

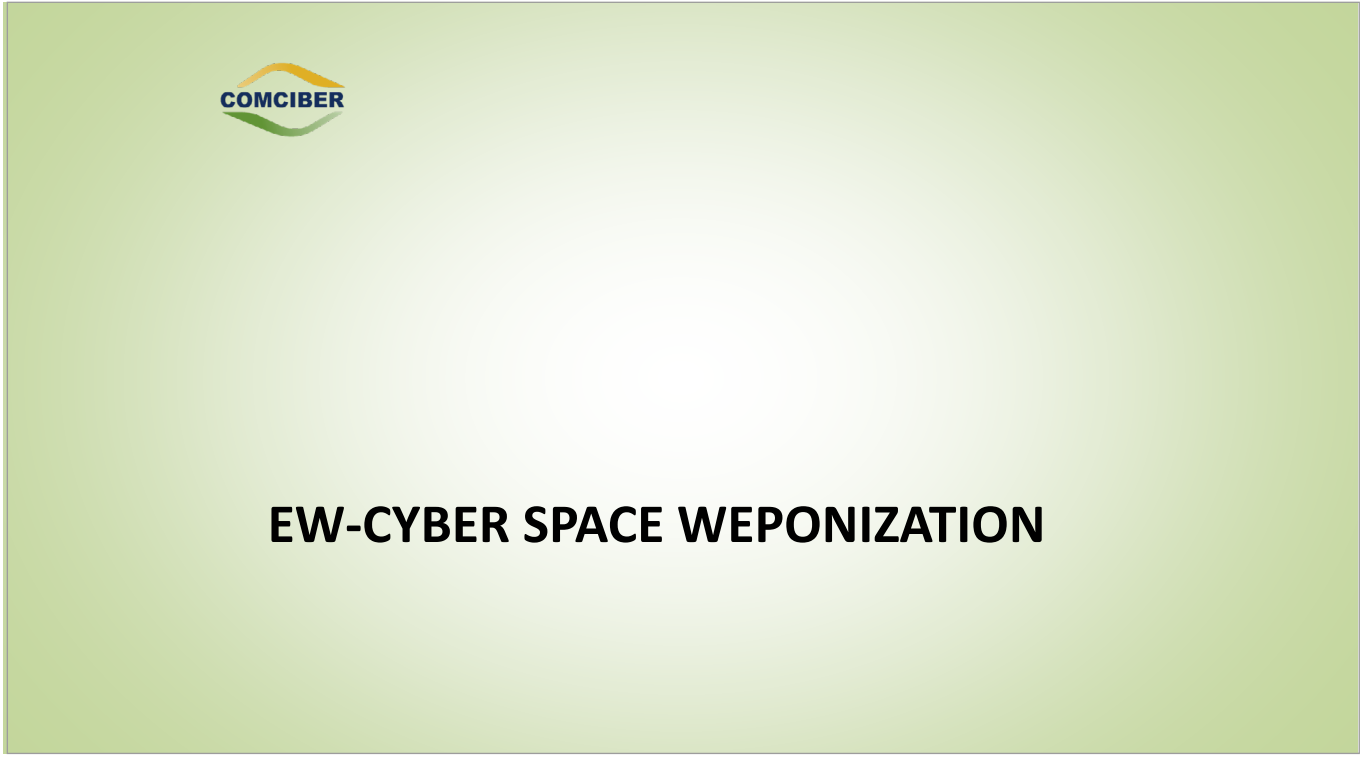
NSA "working" at the Supply Chain

(TS//SI//NF) Such operations involving **supply-chain interdiction** are some of the most productive operations in TAO, because they pre-position access points into hard target networks around the world.

(TS//SI//NF) Left: Intercepted packages are opened carefully; Right: A "load station" implants a beacon

COMCIBER

State objectives can be backed by many partners


COMCIBER
EW-CYBER SPACE WEPONIZATION

The competition for Cyber-EW integration

- Russian case in Syria: In 2017 fleet of 13 UAVs, in action to attack the base used by the Russians, was "disabled" with a mixture of media, including kinetics
  - 6 shot down by anti-aircraft missiles;
  - 7 were landed, speculation is speculated through joint action of jamming and cyber attack
- The U.S. created the Terrestrial Layer Intelligence System in 2017 with the aim of carrying out, among others, the integration of actions and systems and data fusion.
  - Army will put CEMA unit in each brigade (2019)!
- China, Australia, etc. also have initiatives

COMCIBER

MINISTERS NAVY ARMY AIR FORCE DEPARTMENT SEARCH

Australian Government
Department of Defence
Science and Technology

DST Science and Technology for Safeguarding Australia

HOME DISCOVER DST OUR SCIENCE PUBLICATIONS EVENTS PARTNER WITH US CAREERS MEDIA CENTRE CONTACT US ALUMNI

Home › Discover DST › Our research divisions › Cyber and Electronic Warfare Division

Share Print page Larger text Smaller text

DISCOVER DST

- DST at a glance
- About DST
- Our role
- Our value proposition
- Our leadership
- Our corporate divisions
- Our research divisions
  - Maritime Division
  - Land Division

CYBER AND ELECTRONIC WARFARE DIVISION

Cyber and Electronic Warfare Division undertakes research and development focused on identifying, analysing and countering threats to Australia's defence and national security through electronic means.

Cyber and Electronic Warfare Division integrates S&T capabilities to cover the continuum of the cyberspace and electromagnetic environment.

The Division produces and validates concepts, tools and techniques for protecting Australia's Army, Navy, Air Force, Defence Intelligence and broader national security agencies against such threats, and provides expert technical advice to major Defence acquisitions.

Cyber and Electronic Warfare Division integrates science and technology capabilities across cyber, electronic warfare (EW), signals intelligence, and communications to cover the continuum of the cyberspace and electromagnetic environment.

The division applies its capabilities to support situational awareness of the cyber and electromagnetic environment (including through systems, networks, signals and electromagnetic spectrum analysis), reliable and resilient cyber and EW systems (including

KEY INFORMATION

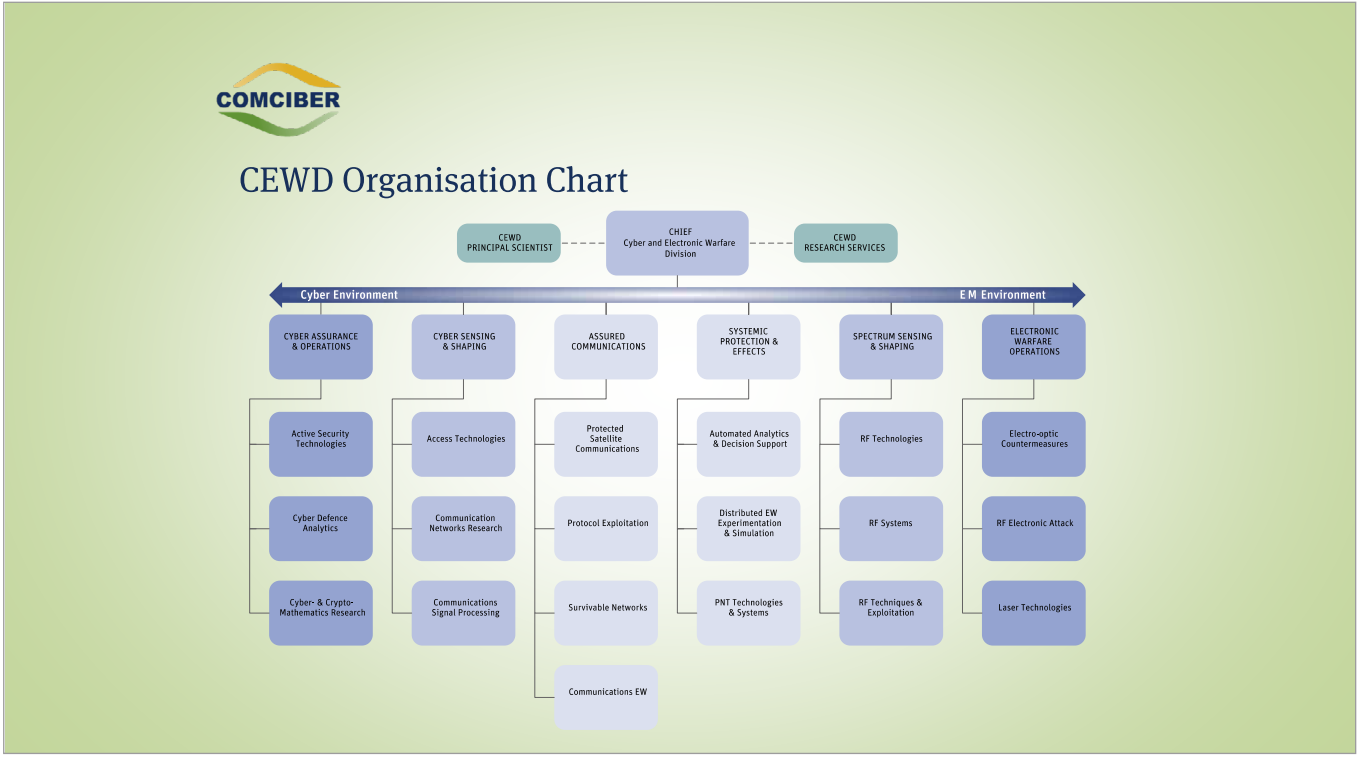
CONTACT
+61 8 7389 5779

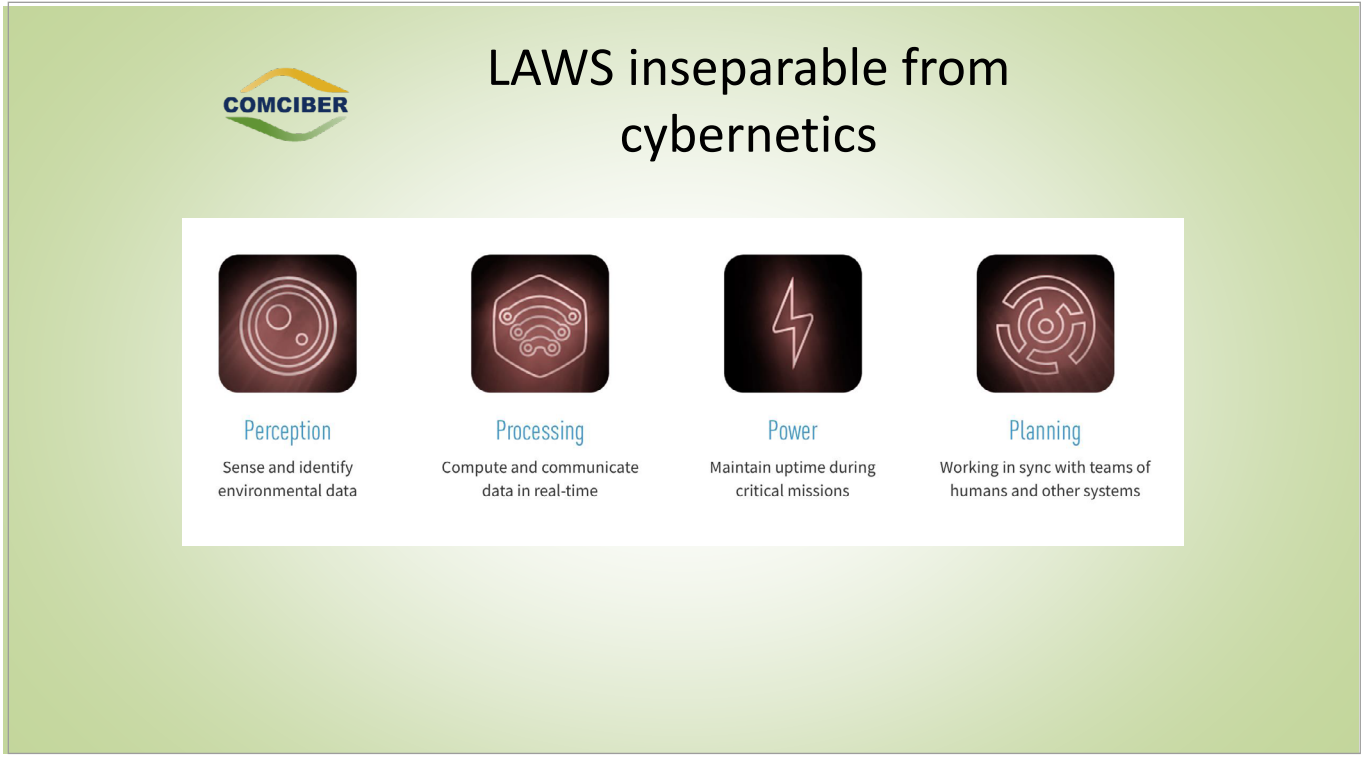
CHIEF OF DIVISION
Dr Dale Lambert (view profile)
CCEWD@dsto.defence.gov.au

LOCATIONS:
DST Group Edinburgh (headquarters)

ATTACHED FILES

Cyber and Electronic Warfare Division Strategic Plan 2016 - 2021
PDF document
(2.03 MB)


COMCIBER
CEWD Organisation Chart
CEWD PRINCIPAL SCIENTIST
CHIEF Cyber and Electronic Warfare Division
CEWD RESEARCH SERVICES
Cyber Environment
EM Environment
CYBER ASSURANCE & OPERATIONS
CYBER SENSING & SHAPING
ASSURED COMMUNICATIONS
SYSTEMIC PROTECTION & EFFECTS
SPECTRUM SENSING & SHAPING
ELECTRONIC WARFARE OPERATIONS
Active Security Technologies
Cyber Defence Analytics
Cyber- & Crypto-Mathematics Research
Access Technologies
Communication Networks Research
Communications Signal Processing
Protected Satellite Communications
Protocol Exploitation
Survivable Networks
Communications EW
Automated Analytics & Decision Support
Distributed EW Experimentation & Simulation
PNT Technologies & Systems
RF Technologies
RF Systems
RF Techniques & Exploitation
Electro-optic Countermeasures
RF Electronic Attack
Laser Technologies


COMCIBER
LAWS inseparable from cybernetics
Perception
Sense and identify environmental data
Processing
Compute and communicate data in real-time
Power
Maintain uptime during critical missions
Planning
Working in sync with teams of humans and other systems

COMCIBER

But cyberweapons usually end-up in the wild...


COMCIBER
Questions
Dr. Roberto Gallo
President ABIMDE <presidencia@abimde.org.br>
CEO Kryptus <gallo@kryptus.com>

Statement by the Chinese Delegation

Chen Yongcan
Deputy Consul General of China

Excellencies, Dear Colleagues,

Ladies and gentlemen,

Good afternoon!

We would like to thank Brazil for hosting the Seminar on LAWS in the wonderful city of Rio. Here I would like to share with you the statement by the Chinese Delegation on the three topics of this seminar.

I. On the issue of human-machine interaction and human control

The fundamental purpose of our discussion on the issues of human-machine interaction and human control is to ensure human intervention at one or various stages of developing and using LAWS,

in order to prevent it from causing indiscriminate effects or being excessively injurious. The essence of the issue is the human-machine relation, including the functional division, means of interaction between humans and machines. Therefore, research should be treated in the framework of human-machine-environment, taking into full consideration the elements like weapons, people, aim, and scenery, instead of simplifying the treatment.

Last year, the GGE on LAWS agreed on a new guiding principle on human-machine interaction, which may take various forms in objective circumstances. We should ensure that the human-machine interaction applied in LAWS complies with applicable international law, in particular IHL. It reflects the common understanding of all parties about human-machine interaction at this stage and is significant in directing the future development of LAWS. All parties can continue the discussion on the basis of the principle.

II. On the appliance of IHL and a new protocol

As a future means or method of warfare, LAWS should be in compliance with international humanitarian laws such as the Geneva Convention of 1949 and the two Additional Protocols in 1977, including the principles of restriction, distinction, and proportionality. However, there are uncertainties in the application of the aforementioned principles. For example, considering the actual level of technological development, can LAWS distinguish civilians from combatants? Can the weapon system make judgments according to the principle of proportionality in the environment of a dynamic battlefield? In addition, if it violates IHL, how can it be held accountable? There are not clear answers to these questions so far. For that reason, it is debatable whether the existing IHL is adequate or not.

The discussion in recent years shows that no parties are willing to develop LAWS completely out of human control. This is an important

basis for our work. Under the precondition of resolving the definition of LAWS, the Chinese side supports formulating a legally binding short protocol, taking as an example Protocol IV on Blinding Laser Weapons of the CCW, in order to restrict and normalize the use of LAWS.

When we discuss the possible options in the future, we should fully consider the complexity of LAWS, by adopting a pragmatic attitude and resolving the issues systematically. Before getting the real results, the Chinese side encourages every state to enforce orientation and supervision over the related technological and industrial development. The 11 Guiding Principles reached by the GGE have active significance and should be used as a reference by all states. The GGE can continue to discuss based on the principles and try to put forward more pragmatic and effective principles.

III. On the strategic and military dimensions of LAWS

As a product of new technological development and military revolution, LAWS is one of the concrete examples of military application of AI technology. Nowadays, the strategic and military dimensions of LAWS draw much international attention, and the humanitarian, legal, and ethical results that it might cause have also raised concern worldwide. However, there are different viewpoints in this regard. Some parties are worried that LAWS will lower the threshold of war by reducing the cost and casualties of war, and thus will increase the possibility and frequency of using force. Some are worried about the challenges to the international security system brought by the results such as arms races, proliferation, and abuse that might be triggered by LAWS. In particular, LAWS may cause disastrous consequences when they fall into the hands of non-state actors and terrorist organizations. Meanwhile, some believe that LAWS have advantages in terms of cost-effectiveness ratios, time of reaction, collateral casualties, and application environment, enabling

them to help or partly substitute human work to effectively avoid humanitarian disasters. Some consider that it is hard to reach a conclusion at this stage, since the technologies applicable to LAWS are complicated and developing very fast, and their future is quite uncertain.

The Chinese side supports the international community in discussing LAWS issues, in order to give an objective and comprehensive assessment of their implications, and then formulate relevant international rules through negotiation, so as to make the best use of the advantages and bypass the disadvantages. Meanwhile, considering that the new technologies applicable to LAWS are dual-use in nature, as they could serve both military and civilian purposes, we should respect the right of peaceful use shared by all states, so as not to hinder scientific and technological development and social progress, let alone set discriminatory technical barriers, using the excuse of non-proliferation to harm the legitimate and equal right of the developing countries to access new technology.

It is significant to establish the GGE on LAWS in the framework of the CCW to discuss the issues of technology, military use, and the application of international laws for such weapons systems. The Chinese side commends the GGE for agreeing on the 11 Guiding Principles, which offer basic guidelines to regulate the development directions of LAWS. We are glad to see that, last year, the CCW High Contracting Parties unanimously adopted the new mandate for the GGE to continue to discuss the relevant aspects of LAWS.

China has always participated actively in the related international discussion and the work of the GGE in a constructive manner, including putting forward Chinese proposals on the characteristics of LAWS. We stand ready to continue the exchange with all parties, in a bid to contribute to reaching more common ground on LAWS.

Thank you!

Autonomy in Weapons Systems and Strategic Stability

Moa Peldán Carlsson and Vincent Boulanin
Stockholm International Peace Research Institute (SIPRI)

The discussion on autonomous weapons systems today revolves mainly around their legal, ethical, and operational considerations. Over time, the implications for strategic stability have been somehow overlooked. This can be explained by the fact that the framework in which the conversation is taking place—the CCW's GGE on LAWS—is primarily concerned with the humanitarian risks posed by the use of conventional weapons in armed conflicts. However, autonomous weapons raise a broader set of challenges. They can be disrupting even *outside* the context of armed conflicts, as their development and potential proliferation could modify the status quo in great power relations, affect states' sense of security and thereby undermine strategic stability and global security.

The GGE process on LAWS could play an important role in mitigating the risks that autonomous weapons systems pose to strategic stability. The purpose of this paper is to explore why and how the GGE on LAWS could give greater consideration to this topic. This paper is drawing from the findings of a SIPRI project entitled "Mapping the Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk."[1] The project is focused on the impact that recent advances in machine learning and autonomy could have on nuclear-armed states' future military modernisation plans and the challenge that these developments could raise as far as strategic stability and nuclear risk reduction are concerned.

Definitions and concepts

First of all, it is useful to clarify what we mean by autonomous weapons systems and strategic stability, since these two concepts may be subject to different interpretations.

There is no internationally agreed definition of autonomous weapons systems; the CCW discussions on how autonomous weapon systems can and should be defined are unresolved. From our perspective, thinking of autonomy as a general feature of weapons systems is technically imprecise and conceptually misleading. Autonomy is better thought of in relation to specific functions or capabilities within a system, be it from a technical, legal, or ethical standpoint.[2] Autonomous navigation capability generates

1 Boulanin, V. (ed.) ***The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk***, vol. I, ***Euro-Atlantic Perspectives*** (SIPRI: Stockholm, May 2019), <https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf>.; Saalman, L. (ed.) ***The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk***, vol. II, ***East Asian Perspectives*** (SIPRI: Stockholm, Oct 2019), <https://www.sipri.org/sites/default/files/2019-10/the_impact_of_artificial_intelligence_on_strategic_stability_and_nuclear_risk_volume_ii.pdf>; Topychkanov, P. (ed.), ***The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk***, vol. III, ***South Asian Perspectives*** (SIPRI: Stockholm, Apr 2020), <https://www.sipri.org/sites/default/files/2020-04/impact_of_ai_on_strategic_stability_and_nuclear_risk_vol_iii_topychkanov_1.pdf>.

2 Boulanin, V. and Verbruggen, M., ***Mapping the Development of Autonomy in Weapon Systems*** (SIPRI:

different technical requirements and challenges than autonomous targeting. Hence, for the purpose of this paper, we would rather talk about "autonomy in weapons systems" than "autonomous weapons systems." This conceptual shift provides us with the opportunity to be more granular in the way we report about recent technological advances—notably artificial intelligence, machine learning and applications thereof—and associate challenges.

The concept of strategic stability was originally coined during the Cold War to describe a situation where both the USA and the Soviet Union would lack incentives to launch a first nuclear strike.[3] Since then, the concept has acquired different meanings. It has been described more broadly as "the absence of armed conflict between nuclear-armed states" and most broadly as the "regional and global security environment in which states enjoy peace and harmonious relations."[4] In this paper, strategic stability is understood in its narrowest and traditional sense as a state of affairs characterized by crisis stability (absence of incentives for any country to launch a first nuclear strike) and arms race stability (absence of incentives to build up nuclear forces). It therefore primarily concerns the relationship between nuclear-armed states—but not only that. The central feature of strategic stability from this standpoint is that nuclear countries are confident that their adversaries, whether nuclear-armed or not, would not be able to undermine their nuclear deterrent capability—i.e. second-strike capability—using nuclear, conventional, or other non-conventional means. Therefore it is where autonomy comes into the picture: advances of autonomy in

Stockholm, Nov. 2017), <https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf>.

3 Steinbruner, J. D., "National security and the concept of strategic stability," ***Journal of Conflict Resolution***, vol. 22, no. 3 (Sep. 1978), pp. 411-28, <https://doi.org/10.1177/002200277802200303>.

4 Edward Warner cited in, Acton, J. "Reclaiming Strategic Stability," ed. E. Colby and M. Gerson ***Strategic Stability: Contending Interpretations***, (US Army War College Press: Carlisle Barracks, 2013), p.117.

weapon systems could both improve and reduce the confidence that nuclear-armed states have in their deterrence capability.

Autonomy and the nuclear deterrence architecture

How could advances of autonomy improve nuclear-armed states' confidence in their nuclear deterrence capability?

To understand the connection between autonomy and the nuclear deterrence architecture, it is useful to make a distinction between two types of autonomy: autonomy *at rest*, and autonomy *in motion*.[5] Autonomy at rest refers to applications that operate virtually in software; these include various types of planning and decision support systems, but also cyber security systems. Autonomy in motion refers to applications that allow systems to have a presence in, and act on the physical world—e.g. autonomous navigation and automatic target recognition. In the framework of nuclear deterrence architecture, autonomy at rest would concern areas such as early warning, ISR (Intelligence, Surveillance, and Reconnaissance) data processing and command and control, while autonomy in motion would be relevant for ISR data collection and force delivery. Advances of autonomy could, in other words, find applications in nearly all critical areas of nuclear deterrence.

From a technical standpoint, the current and foreseeable advances of autonomy will derive to a large extent from the progress of machine learning. Machine learning is an approach to software development that means first building systems that can learn and then teaching them what to do using a variety of methods and a lot of data. Machine learning has been around since the beginning of AI research, but has experienced a breakthrough over the last course of the past decade. Machine learning is particularly good at

5 Defense Science Board, ***Summer study on autonomy***, (Final report: US Department of Defense: Washington, Jun. 2016), p. 5.

finding connections between data, which makes it a powerful tool for automating any tasks that require advanced pattern recognition. From this standpoint, the possibilities that machine learning offers in the nuclear realm are wide-ranging. With regard to autonomy at rest, machine learning could be leveraged to boost detection capabilities of early warning systems, improve analyses of ISR data, and enhance the protection of the command and control architecture against cyberattacks. Concerning autonomy in motion, machine learning could be leveraged to enhance the autonomous navigation capabilities of any type of vehicle. In practice, that means an improved possibility for remote sensing operations and force delivery: autonomous remote sensing systems would be able to travel more stealthily than their remote-controlled counterparts and in areas that are hardly accessible for manned and remotely-controlled systems, such as in the deep sea. Delivery platforms, on the other hand—be they conventional or nuclear—could be enhanced with navigation control that could allow them to have manoeuvrability.[6] Further, machine learning could also be used to boost the automated target recognition capability or air- and missile-defence systems.[7]

All of the aforementioned possibilities, if technically realised and adopted, could theoretically improve nuclear-armed states' confidence in their nuclear deterrence capability, as they hold the promise of making them more prepared and responsive to nuclear-

6 Saalman, L. "Integration of neural networks into hypersonic glide vehicles, " inSaalman, L. (ed.), ***The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk***, vol. II, ***East Asian Perspectives*** (SIPRI: Stockholm, Oct. 2019) <https://www.sipri.org/sites/default/files/2019-10/the_impact_of_artificial_intelligence_on_strategic_stability_and_nuclear_risk_volume_ii.pdf>.

7 Boulanin, V. "The future of machine learning and autonomy in nuclear weapon systems," in Boulanin, V. (ed.), ***The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk***, vol. I, ***Euro-Atlantic Perspectives*** (SIPRI: Stockholm, May. 2019) <https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf>; Bronk, J. "The impact of unmanned combat aerial vehicles on strategic stability, "in Boulanin, V. (ed.), ***The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk***, vol. I, Euro-Atlantic Perspectives (SIPRI: Stockholm, May. 2019) <https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf>.

related threats. However, the very same capability that can increase one state's sense of confidence can also be a source of insecurity for another state, and thereby undermine strategic stability.

Destabilising effects on strategic stability

How could advances in autonomy undermine nuclear-armed states' confidence in their own nuclear deterrence capability?

The central source of insecurity is that an adversary's advances in autonomy could make it more capable of threatening one's second-strike capabilities. The fear is that the development of machine-learning-boosted ISR systems and autonomous remote sensing platforms could make one's nuclear force harder to hide, but also to protect.

One particular challenge in this regard is the fact that advances in autonomy hold the promise to make conventional weapons systems more capable and potentially more threatening to nuclear assets. In other words, they could lead to greater "entanglement" between the conventional and the nuclear arena.[8] The perception that one's nuclear capability could be defeated by the conventional means of an adversary could lead some actors to adopt postures or measures that could undermine strategic stability and *in fine* increase the risk of nuclear weapon use.

In terms of posture, one possibility is that a nuclear-armed state would (further) open up to the possibility of using nuclear weapons in response to a conventional attack. In fact, there is an observable increasing political willingness to use nuclear means to retaliate against non-nuclear attacks. For instance, the US and Russia have

8 Entanglement refers to the increasingly intertwined nuclear and non-nuclear systems by increased capabilities in conventional weapon systems. Arbatov, A. et al., "The Escalation through Entanglement How the Vulnerability of Command-and-Control Systems Raises the Risks of an Inadvertent Nuclear War 56," ***International Security*** vol. 43, no. 1 (2018), p. 56-99.

developed positions saying that, in the future, they would possibly respond to conventional attacks with nuclear means.[9]

In terms of measures, there is, first of all, the risk that nuclear-armed states would feel the need to enter into an arms- or capability race. Such a race could be destabilising as it could lead states to adopt the latest AI technologies prematurely or irresponsibly out of fear of lagging behind others. Another destabilising prospect could be that some actors would try to offset the technological disadvantage they have in the field of AI and conventional weapons with further investments in nuclear arsenals. That possibility has already been discussed in Russia as a reaction to the USA's AI-focused Third Offset Strategy.[10] Another concerning prospect is the possibility that some states would renounce to their "no first use" policies or increase their alert statuses for nuclear assets, meaning increasing their readiness to launch a nuclear strike. Perhaps, the most destabilising measure would be that one state would feel it necessary to automate part of its nuclear command and control to deter its adversary with the possibility of an automated retaliation. It is currently unlikely that any of the nuclear-armed states would do so, since the consequences of failure within such a system would be disastrous, however such possibility cannot be entirely excluded given it was explored by the Soviet Union during the Cold War.[11]

One critical point to note is that advances in autonomy do not need to *actually* be realised to become a concern and trigger destabilising reactions.[12] In the realm of strategy, perception of

9 Altmann, J. and Sauer, F., "Autonomous Weapon Systems and Strategic Stability," ***Survival*** vol. 59, no. 5 (2017), p. 117-42.

10 Kashin, V. and Raska, M., Countering the US Third Offset Strategy: Russian Perspectives, Responses and Challenges, S. Rajaratnam School of International Studies (RSIS) Policy Report (RSIS: Singapore, Jan. 2017).

11 Hoffman, D. E., ***The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*** (Anchor Books: New York, 2009).

12 Geist, E. and Lohn, A. J., How Might Artificial Intelligence Affect the Risk of Nuclear War? (Rand

capabilities matters as much, if not more than the capabilities themselves. Destabilising measures could be introduced only based on the belief that that advances in autonomy could offer one's adversary credible options to threaten the survivability and reliability of one's nuclear deterrent.

What role could the GGE on LAWS play for mitigating the risk of undermining strategic stability?

In sum, advances of autonomy, even if strictly contained to the conventional arena, could have an impact on nuclear deterrence relations and in fine strategic stability. The question is then, should and could the GGE process on LAWS do something about it?

It is beyond dispute that strategic stability considerations are not within the mandate of the CCW GGE on LAWS. The CCW is meant to focus on humanitarian consideration, and for that reason does not provide the appropriate forum to discuss in depth the challenges argued above. However, as a current focal point in the international debate on the employment of military use of autonomy, as well as AI and machine learning more generally, there are various ways in which the work conducted by the CCW could reduce the perception of problems from which destabilizing dynamics could emerge.

First, the CCW process could play a critical role in reducing misconceptions about the state of technology. It could reduce the danger of states over-estimating each other's capabilities or the state of technologies and making ill-advised nuclear policy decisions accordingly.

Corporation: Santa Monica, CA, 2018); Rickli, J."The destabilizing prospects of artificial intelligence for nuclear strategy, deterrence and stability," in Boulanin, V. (ed.), ***The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk***, vol. I, ***Euro-Atlantic Perspectives*** (SIPRI: Stockholm, May. 2019) <https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf>.

Second, the CCW's process could generate more transparency around what safety and reliability standards countries adopt in relation to AI and autonomous systems. This could slow down the speed and reduce the risk of immature employment of AI technology.

Third, the CCW's deliberation on the questions of human control could provide identifiable limits in terms of requirements for responsible use of AI technology in nuclear command and control.

In a nutshell, there are various significant ways in which the CCW could help mitigate the risk that autonomy in weapons systems poses to strategic instability, while considering the humanitarian risk of lethal autonomous weapons systems.


Autonomous weapon systems and the impact on strategic stability
Moa Peldán Carlsson
20 February 2020
sipri

Impact on strategic stability

Discussions focus on legal, ethical and operational implications

What about strategic stability???

Does autonomy in weapon systems
... undermine strategic stability?
... lead to an arms race and escalation?
... increase the risk of a nuclear launch?

SIPRI mapping study

- 3 reports, last one out in 2020
- Find on SIPRI's website

Definitions: AWS

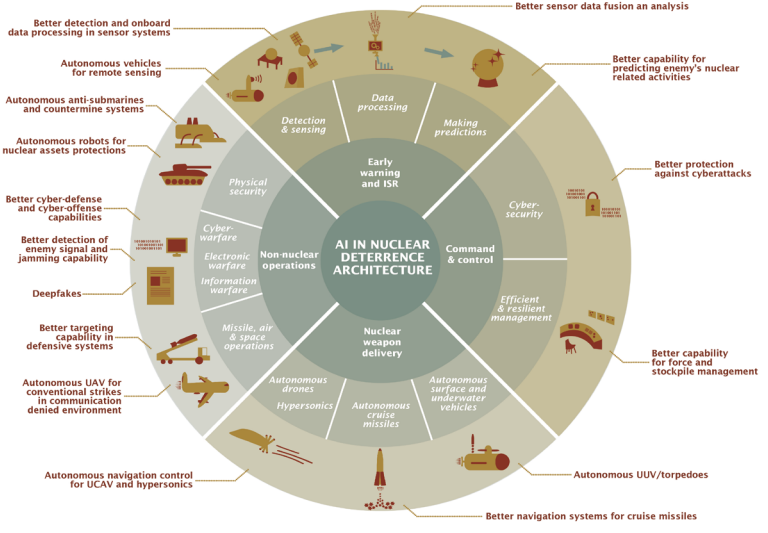
Autonomous weapon systems

Autonomy *in* weapon systems

Definitions: strategic stability

- ”A state of affairs in which countries are confident that their **adversaries would not be able to undermine their nuclear deterrent capability** using nuclear, conventional or other non-conventional means”
- Achieved by mutually assured destruction (MAD)
- Depends on
  1) the possession of second-strike capability
  2) that the capability are credible, effective and survivable

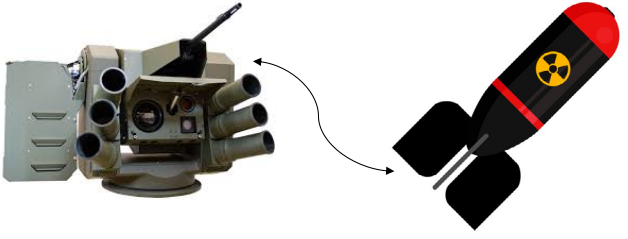
Application of autonomy in nuclear deterrence architecture



sipri 20/02/2020

Entanglement

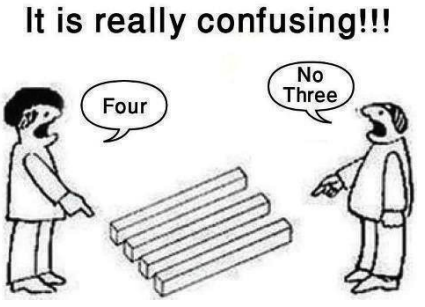
- Connection between conventional and nuclear arms
- Autonomy in weapon systems driver
- Easier to hold nuclear assets at risk
- Willingness to counter conventional attacks with nuclear



sipri 20/02/2020

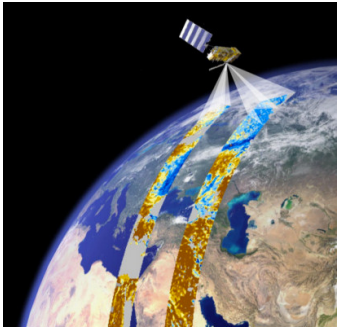
Destabilising effects on strategic stability

- Reinforce asymmetry between states
- Incentives to respond with destabilising measures
  - arms race
  - modernize nuclear arsenals
  - renounce NFU policy
  - increase alert status
  - automate launch
- **Issue of perception**
- Speed of warfare
- Lower threshold of war



Escalation

- Accidental escalation
  - less time for decision-making
  - brittleness in systems
  - operation unknown
- Inadvertent escalation
  - issue of perception
- Deliberate escalation
  - AI generated information



(Stabilising effects on strategic stability)

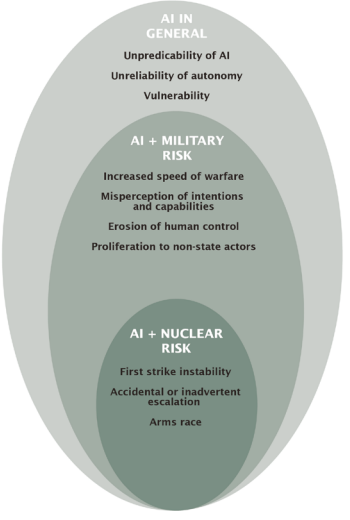
- Mutual vulnerability
- Better prepared to deal with crisis

How serious is the risk?

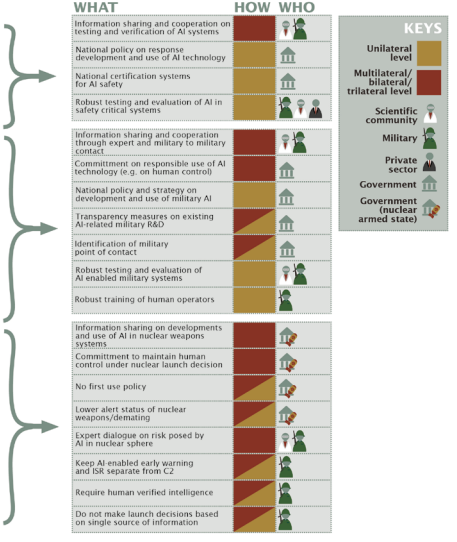
- Autonomy in weapon systems alone is not enough to trigger escalation
- Other key factors

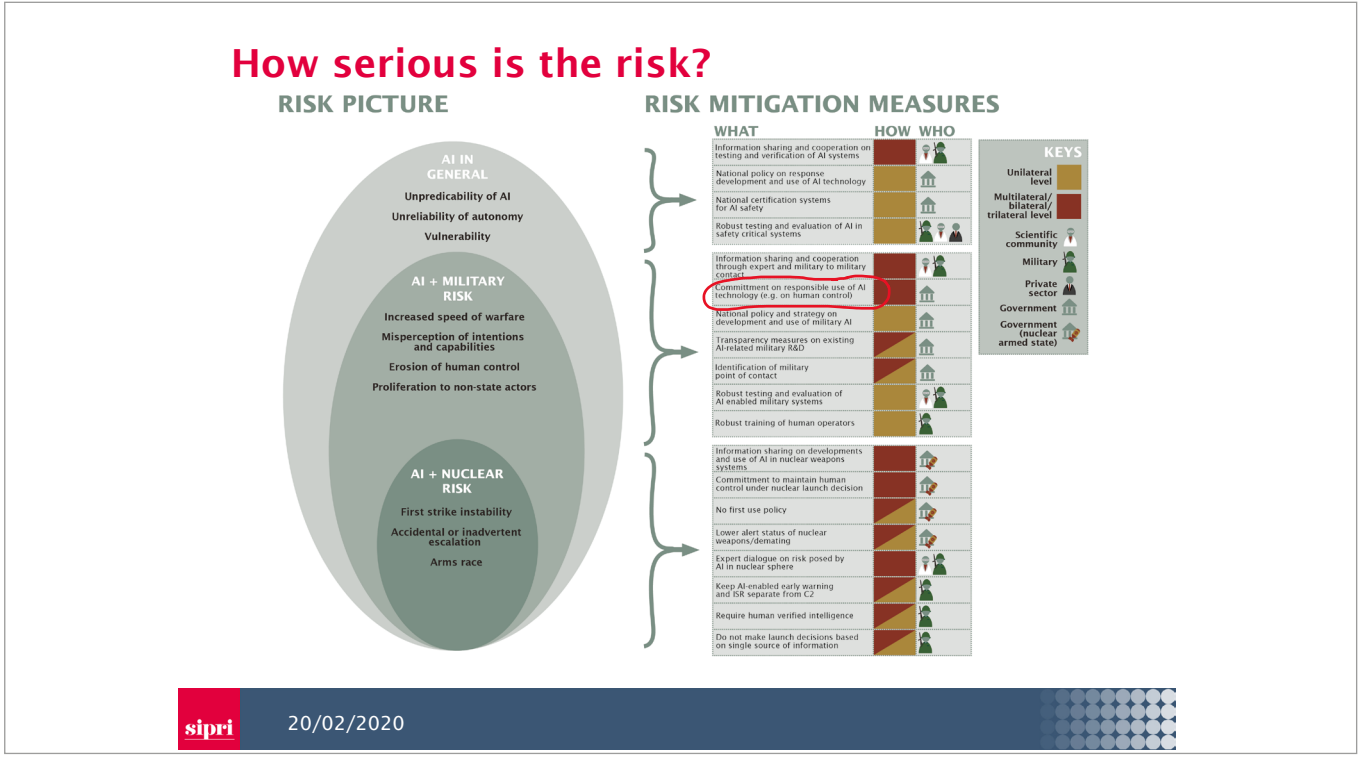
How serious is the risk?

RISK PICTURE



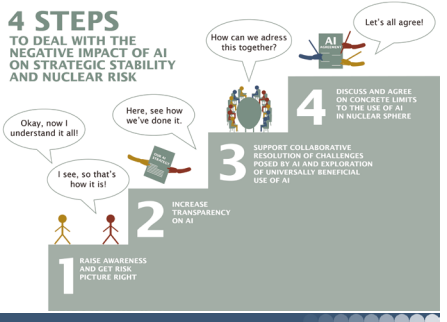
RISK MITIGATION MEASURES




How serious is the risk?
RISK PICTURE
AI IN GENERAL
Unpredictability of AI
Unreliability of autonomy
Vulnerability
AI + MILITARY RISK
Increased speed of warfare
Misperception of intentions and capabilities
Erosion of human control
Proliferation to non-state actors
AI + NUCLEAR RISK
First strike instability
Accidental or inadvertent escalation
Arms race
RISK MITIGATION MEASURES
WHAT
HOW
WHO
Information sharing and cooperation on testing and verification of AI systems
National policy on response development and use of AI technology
National certification systems for AI safety
Robust testing and evaluation of AI in safety critical systems
Information sharing and cooperation through expert and military to military contact
Committment on responsible use of AI technology (e.g. on human control)
National policy and strategy on development and use of military AI
Transparency measures on existing AI-related military R&D
Identification of military point of contact
Robust testing and evaluation of AI enabled military systems
Robust training of human operators
Information sharing on developments and use of AI in nuclear weapons systems
Committment to maintain human control under nuclear launch decision
No first use policy
Lower alert status of nuclear weapons/demating
Expert dialogue on risk posed by AI in nuclear sphere
Keep AI-enabled early warning and ISR separate from C2
Require human verified intelligence
Do not make launch decisions based on single source of information
KEYS
Unilateral level
Multilateral/ bilateral/ trilateral level
Scientific community
Military
Private sector
Government
Government (nuclear armed state)
sipri
20/02/2020

Key takeaways

- Autonomy in weapon systems a driver for entanglement
- Perception can be destabilising
- Key to agree on regulations + share information



Thank you!

Website: www.sipri.org

Email: moa.peldan@sipri.org

FUNDAÇÃO
ALEXANDRE
DE GUSMÃO

Follow our social media

@funagbrasil

The Rio Seminar on Autonomous Weapons Systems, held in Rio de Janeiro at the Naval War College on February 20, 2020, aimed at contributing to the debate on the governance of emerging technologies in LAWS (Lethal Autonomous Weapons Systems) under international law, including IHL (International Humanitarian Law).

The Rio Seminar took place in the framework of the GGE-LAWS of the CCW (Group of Governmental Experts of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons).

Its purpose was to foster discussions among the main participants of the LAWS negotiations—government representatives, international organizations, International Committee of the Red Cross, non-governmental organizations, private sector, and academia—in a multi-stakeholder approach considering its diplomatic, legal, technological, corporate, strategic, and military dimensions. The informal setting enabled a dynamic knowledge sharing, which may help governments and non-governmental delegations in preparing for the GGE activities in 2020, and its recommendations to the next Meeting of the High Contracting Parties, in 2020, and the Sixth Review Conference of the CCW, in 2021.

The video presentations of the Rio Seminar are available at: <https://m.youtube.com/playlist?list=PLY4MsNDouGfge7-IAdRZtdJk2mJrwljVz>.

FUNDAÇÃO ALEXANDRE DE GUSMÃO
www.funag.gov.br

ISBN 978-65-87083-30-8