EXECUTIVE SUMMARY

ARTIFICIAL INELLIGENCE AND AUTONOMOUS FORCE APPLICATION SYSTEMS

Published on September 2021

By Mg. Luciana Micha, Dr Vanina Martínez,
commodore Pablo Farías and Dr Ricardo O. Rodríguez

The present interdisciplinary approach aims to generate concepts that facilitate a comprehensive approach on the risks and threats presented by Autonomous Weapon Systems (AWS), with the purpose of providing strategies that contribute to the current debate on the need of its regulation and/or prohibition.

This paper contributes to the design and development of a set of matrices **that link the main dimensions and processes associated with AWS** and the guidelines derived from the principles and spirit of International Humanitarian Law, Human Rights (HR) and the guiding principles for a reliable and responsible Artificial Intelligence (AI).

As specific products, the matrices, which apply independently from the specific AI models that underlay AWS, **focus on autonomy** (human-machine interaction) **and the principles of reliability that intrinsically must be verified.**

At the same time, these analyses answer the key questions raised within the United Nations in the Group of Governmental Experts (GGE) of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons, where the matrices attempt to articulate themselves to international and regional dialogues and provide responses from IHL, AI and the key dimensions analyzed: **Decision Making Levels, AWS' Tasks and Functions, Operational Employment Domains and environmental requirements, target types and AWS' lethal capacity**, focusing on the Human System **interaction required in each case.**

The analysis dynamic developed presents a different approach and perspective to the one that is now being discussed, which is limited since it places as center of gravity the considerations regarding use, regulation and prohibition focused on a product ("machine or robot") and assessments arising from their purely tactical use (engagement).

By understanding the complexity and challenges of autonomous force application systems and subsystems, this proposal broadens the field of conceptual analysis to provide depth to the analysis of AWS, systems and processes, and their impact throughout the decision-making cycle (Political-Operational-Tactical).

It is acknowledged that meaningful human control over the critical functions of AWS in relation to the process of command and control, targeting and use of force is imperative. Once the sequencing of the decision process has been determined, **the degree of significant human control must be set through IHM (addressed in Matrix 1),** followed by considering **the requirements of the operational environment and characteristics of the objective's context (Matrices 2.1, 2.2 and 3)**. Finally, the requirements for AWS are determined in relation to **the principles of reliability for AI (matrix M4)** in order to guarantee the degree of significant human control previously determined.

Through a deep multidimensional analysis, this document allows to identify **preventive and/or anticipative actions/strategies**, besides the reactive ones. This sheds light to objectives and intermediate (sub-optimal) strategies which contribute both to the development judging elements from a comprehensive approximation as its contribution to the continuity of dialogue instances and the progressive construction of consensus for its use, regulation, non-proliferation and/or prohibition.

With regards to the Key Principles of IHL, through a detailed analysis on the compliance with the rules of IHL in relation to AWS, the critical principles of **Humanity**, **Reliability and Limitation** are taken as being the most relevant when trying to establish a limit between legality and illegality in the use of AWS.

In relation to the rest of the principles of IHL, as long as AWS cannot demonstrate empirically or theoretically the ability to effectively distinguish between combatants and non-combatants, Principle of Distinction, and that its actions allow the application of the principles of Proportionality and Precaution, its use should be preventively prohibited.

Regarding the **"Principle of Humanity",** considered as one of the main sources of international law in general and IHL in particular, it is analyzed how this principle contradicts AWS due to the lack of guarantees that these systems have in the decision-making process in relation to autonomous action and the need to limit the effects of armed violence on the safety and health of people and civilian property.

In this sense, the application of the **Martens Clause** shows that AWS would not act under the precepts and mandates of humanity and public conscience due to a lack of human emotions, empathy, compassion and ethical principles which belong to humanity.

Therefore, the principle of **humanity** not only carries the effects and consequences of the use of AWS. But also, for the principle's analysis and application, the system itself and its associated decision-making processes must be taken into consideration. The principle of **humanity must be present both in the means, the systems to be employed and their consequences**, understanding that there must be a **significant human control in any weapon system**, regardless of the technological advance that exists, especially in the case of anti-personnel weapons.

With regards to the Principles of Reliability and Limitation, it is pointed out that, in view of the mandatory nature of Article 36 of Additional Protocol I, effective human control is conclusive when attributing personal **responsibility** throughout the stages of design, development and use. Thus, it is States, parties and individuals involved in armed conflict who are compelled by IHL, but not systems and machines.

Article 36 of the Additional Protocol to the Geneva Conventions states that "the right of the parties of a conflict to choose methods or means of warfare is not unlimited and the use of weapons that cause superfluous injury or unnecessary suffering is prohibited". The document emphasizes that **what is not expressly prohibited or restricted, NOT PERMITTED, should be evaluated under the general rules of IHL.**

For this reason, and considering the existence of the Guide for the Legal Review of New Weapons prepared by the International Committee of the Red Cross (ICRC) that takes into consideration the measures for the implementation of Art 36 establishes that **"in the absence of a specific prohibition or restriction, the assessment of the legality of a new weapon system should be carried out in the light of general prohibitions and in accordance with the basic principles of IHL and customary law".**

In this way, States are urged to establish internationally agreed limits to ensure the protection of civilians by complying with IHL in the face of the dangers that these systems represent if they are not sufficiently

predictable, understandable and explicable.

Therefore, the ICRC expressly settles the need of a **total ban on systems designed for the use of force against people and the regulation of AWS' design and use**.

The work also stresses the need for a review of **the current list of Required Empirical Data within the Guide for the Legal Review of New Weapons**, where the empirical data that would currently determine the legality or illegality of the new weapons systems is: 1) Technical description of the weapon, 2) Technical functioning of the weapon, 3) Health considerations and 4) Environmental considerations.

It is acknowledged that given the advancement of technology and the application of AI to weapons, other important factors should be considered within this list in order to **measure their performance and ensure the specific technical reliability** of AWS. Therefore, it should be added to the ICRC list the following considerations about a **5th element of technical and operational reliability (trustworthiness) which implies that it can be justifiably demonstrated - theoretically and operationally - that the system/s are foreseeable, with significant human control and allocation of individual responsibilities throughout the design, development, use and final disposition process.**

In this sense, this 5th element would oblige to review and anticipate the inherent and associated risks of AWS, which include among others: 1) Design, programming and production problems, 2) Possibility of biases in algorithms (algorithmic bias) due to human-specific subjectivities (or groups) about personal /cultural /ethical appraisals. Such biases would be an aggravating factor in weapons targeting humans, 3) Inconsistencies or hidden defects, 4) Incomplete, ambiguous, contradictory, irrelevant and/or excessive data and information 5) The possibility of hacking, primarily those working in nodal or networked systems, 6) Deception or interference with sensors and/or sources. Adverse actions (adverse actions) or interference to sensors, 7) Challenges in identifying action traceability in order to place accountability and responsibility accounts where the use of autonomy and the challenges of the black box are a major obstacle, 8) Operating standards and targeting not covered by IHL 9) Modification of its performance in comparison to the design and behavior conditions, as well as the inability to solve complex situations that require an adequate situational awareness. 10) Associated considerations on space, time, among others.

The challenge is to identify throughout the process, from decision and implementation levels, an appropriate balance between limitation, risk and acceptability and how to transform this inherent risk (uncertainty) in a residual risk scenario that makes its use in the field of IHL application politically and militarily acceptable, which at present is **impossible due to its immature development and in the future, the limitations foreseen by the principles of Humanity, Limitation or Liability.**

With the considerations of analyzed IHL, the conceptual framework provided by the authors focuses on different conceptual matrices that would make it possible to evaluate a system's performance, identifying its extreme cases and failure points, to model the ways in which it would fail and its possible effects.

Conceptual Matrices

Matrix 1 links **AWS' Functions/tasks with Decision Levels**. This matrix aims to preventively focus IHM on the different levels of the military political decision-making process that provides an adequate IHM according to the principles and obligations of IHL, identifying the life cycle's phase in which human control should be exercised.

Matrix 2 attempts to address the following question: **How do elements of the operational environment-influence the quality and scope of IHM?** Therefore, Matrix 2 allows, in general, from the framework of the **Operational Domain on the Use of AWS**, to link the requirements of the **operational environment with the levels of force application** (non-lethal/ lethal capacity) of AWS. Cells indicate the type of human interaction/devices (IHM).

Given the existence of different domains, for their conceptual approach, the matrix is presented in **Matrix 2.1 and Matrix 2.2:**

On the one hand, **Matrix 2.1:** addresses the **Operational Domain in general**, where the requirements the operational environment/application's variables of non-lethal/lethal force are analyzed. It establishes the minimum acceptable degree of human control at the time of force application.

On the other hand, **Matrix 2.2** focuses on the **Aerospace Domain**, as an example of application for a particular Operational Domain - and addresses the demands of the different variables of the operational environment versus the application of -non-lethal/ lethal force. It is responsible for identifying the minimum acceptable degree of human control for the application of force in this context.

For its part, **Matrix 3** is responsible for **determining the degree of quality and scope that the IHM should have in relation to the levels of force application, the objectives' characteristics and different stages of conflict management to remain within the framework of established IHL and ROEs.** That is why it links the Objectives (target types and conditions depending on the strategic situation) and the Non-lethal/ Lethal capacity of AWS. It establishes the minimum acceptable human control in the system and the interaction between the human and the device in relation to the objectives set and its lethal force capability.

Finally, Matrix 4 attempts to address the question: **What guiding principles of AI must AWS**

have according to the different stages of the force application process to ensure adequate meaningful human control? It therefore focuses on establishing the AI guiding principles in AWS in relation to the different stages of the force application process in order to ensure meaningful human control. These AI guiding principles that AWS must comply with are: Predictability, Understanding, Justification/Accountability and Explanation.

Methodologically, matrices should not be seen as independent from each other. AWS should be analyzed in their design and usage context through all the above-mentioned matrices to verify their possible development, application, or need for limitation, regulation and/or prohibition. At the same time, the door is left open for the creation of more matrices that contribute to a more detailed analysis of each system/operating environment.

Conclusion

In conclusion, this document presents a concrete methodology for the conceptualization and characterization of AWS that serves as a starting point to analyze the demands of the minimum Human Interaction System required and a proper assessment of these systems' technical and operational reliability in order to anticipate inherent and associated risks.

The set of conceptual matrices presented would help to evaluate the performance of a system with the characteristics described and will provide the analysis of acceptable degrees of rationality and predictability in the development processes of this technology.

This occurs as a result of the urgency to leave the "Moving Target" dynamic of AI, which is constantly evolving, causing that the proposed standards become obsolete or outdated in face of the continuous emergence of new risks and challenges that are generated in relation to AI for the future.

As long as this scenario does not take place, it will be necessary for States to take actions aimed at the containment, limitation and prohibition of those key aspects of the process that undermine compliance with the principles of International Humanitarian Law (IHL), International Human Rights Law (HR) and the guiding principles for a reliable and responsible AI.

Finally, we understand that based on the transcendent risks, threats and impacts generated by AI - in general - and AWS - in particular - in the future of humanity (ethical dimension), the management of conflict (political-strategic dimension) and the engagement (operational-tactical dimension), nor the most advanced technology development can replace those aspects of the human nature that allows it to avoid, overcome and/or interrupt the causal chain of a conflict/violence dynamic.

For this reason, the principle of humanity will constitute the basic universal criteria of application to the conception, design, development, certification and use of AWS, assuring that humans -in a permanent way- are the stakeholders of their reality, creation and future.

Sources and Bibliography:

Amoroso, D. and Tamburrini, G., "Toward a Normative Model of Meaningful Human Control over Weapons Systems" Available on 22/09/2021 in File:///C:/Users/administrador/Downloads/toward-a-normative-model-of-meaningful-human-control-over-weapons-systems.pdf

CCW (2019), "Guiding Principles Affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems" Annex available on 21/09/2021 in https://www.ccdcoe.org/uploads/2020/02/UN-191213_CCW-MSP-Final-report-Annex-III_Guiding-Principles-affirmed-by-GGE.pdf

Comisión Europea (2019), "Directrices para una IA fiable". Available on 22/09/2021 in bit.ly/2RYbh8D

Coupland, R. (2001), "Humanity: What is it and how does it influence international law?", ICRC, available on 21/09/2021, https://www.icrc.org/en/doc/assets/files/other/irrc-844-coupland.pdf

Davidson, Neil (2017) "A legal perspective AWS under IHL" UNODA OCCASIONAL PAPERS N°30

Díaz, M. y Muñoz, W. (2020) "Los riesgos de las armas autónomas letales: una perspectiva interseccional latinoamericana". Red de Seguridad Humana para América Latina y el Caribe. Available on 22/09/2021 in https://img1.wsimg.com/blobby/go/98c6dc90-096f-4389-9309-f1a33c0cad73/downloads/Los%20riesgos%20de%20las%20armas%20aut%C3%B3nomas%20una%20perspec.pdf?ver=1623789686395

Dignum, Virginia (2019) "Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way" in "Artificial Intelligence: Foundations, Theory, and Algorithms", ISBN 978-3-030-30370-9, pp. 1-120.

Ejercito argentino (1999) “Reglamento de conducción para el instrumento militar terrestre” Public military document.

In depth interviews to Brigadier (R) Armando Bonadío, director of the Institute of Aeronautical and Space Law, director and professor of the INDAE IHL Course (2021)

G. Persi Paoli, A. Spazian, A. Anand, (2021) “Table-Top Exercises on the Human Element and Autonomous Weapons Systems: Summary Report”, Geneva, Switzerland: UNIDIR.

Gariglio, Damián (2020) “Los Sistemas de Armas Autónomos Letales y el Desbalance de Poder”. Red de Seguridad Humana para América Latina y el Caribe y Centro de Estudios de Política Internacional de la UBA, 2020. Available on 22/09/2021 in https://18df0113-5ebe-4413-800c-0bfd2e8147bb.filesusr.com/ugd/7cccdc_b388e2ca492b4f7c841d6eff14a3a50f.pdf

Holland Michel, Arthur, (2021). “Known Unknowns: Data Issues and Military Autonomous Systems”. Geneva, UNIDIR. Available on 22/09/2021 in https://unidir.org/known-unknowns

Holland Michel, Arthur. (2020). ‘The Black Box, Unlocked: Predictability and Understandability in Military AI.’ Geneva, Switzerland: United Nations Institute for Disarmament Research. doi: 10.37559/SecTec/20/AI1

ICRC (2021) “Position on Autonomous Weapons Systems”. Available on 21/09/2021 in https://www.icrc.org/en/download/file/166330/icrc_position_on_aws_and_background_paper.pdf

ICRC (2021B)” Armas autónomas: el CICR recomienda adoptar nuevas normas” [Declaración]. Available on 22/09/2021 in https://www.icrc.org/en/document/autonomous-weapons-icrc-recommends-new-rul

INTERNATIONAL COURT OF JUSTICE (1196) “Advisory Opinion on the Legality of the Threat or Use of Nuclear Weapons” Available on 21/09/2021 in https://www.icj-cij.org/public/files/case-related/95/095-19960708-ADV-01-00-EN.pdf

Lawand Kathleen, Coupland Robin y Herby Peter (2006), “Guía para el examen jurídico de las armas, los medios y los métodos de guerra nuevos.” ICRC, available on 21/09/2021 in https://www.icrc.org/es/doc/assets/files/other/icrc_003_0902.pdf

López Díaz, P. (2009). “Principios fundamentales del Derecho Internacional Humanitario”. REVISMAR. pág. 230 a 238. Available on 22/09/2021 in https://revistamarina.cl/revistas/2009/3/lopez.pdf

Martínez Vanina y Rodríguez Ricardo (2020) “Aportes al debate del uso de la IA para aplicaciones armamentísticas” en Red de Seguridad Humana para América Latina y el Caribe”. Available on 22/09/2021 in https://img1.wsimg.com/blobby/go/98c6dc90-096f-4389-9309-f1a33c0cad73/downloads/Contributions%20to%20the%20debate%20on%20the%20use%20of%20Arti.pdf?ver=1630078276981

Matthias, Andreas (2004), "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata" en Ethics and Information Technology

Micha, L. y Farías, P. (2021) "La evolución de tecnologías disruptivas y los sistemas de armas autónomas letales: consideraciones desde el ámbito militar". Red de Seguridad Humana para América Latina y el Caribe y Centro de Estudios de Política Internacional de la UBA, (CEPI) marzo, 2021. Available on 22/09/2021 in https://img1.wsimg.com/blobby/go/98c6dc90-096f-4389-9309-f1a33c0cad73/downloads/La%20evoluci%C3%B3n%20de%20tecnolog%C3%ADas%20disruptivas%20Los%20SA.pdf?ver=1627396384155

Muggleton, S.H., Schmid, U., Zeller, C. (2018) "Ultra-Strong Machine Learning: comprehensibility of programs learned with ILP". Mach Learn 107, 1119–1140 Available on 22/09/2021 in https://doi.org/10.1007/s10994-018-5707-3

Sparrow, Robert (2007) "Killer Robots" en Journal of Applied Philosophy, pp 62-77

UNIDIR (2019) "El elemento humano en las decisiones sobre el uso de la fuerza" Available on 22/09/2021 in https://www.unidir.org/publication/el-elemento-humano-en-las-decisiones-sobre-el-uso-de-la-fuerzaG

UNIDIR (2018). "Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies" Available on 22/09/2021 in https://unidir.org/publication/algorithmic-bias-and-weaponization-increasingly -autonomous-technologies

Virginia, Dignum (2019) "Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way". Artificial Intelligence: Foundations, Theory, and Algorithms, Springer, ISBN 978-3-030-30370-9, pp. 1-120.

About the authors

Mg. Luciana **Micha**

commodore Pablo **Farías**

Dra. Vanina **Martinez**

Dr. Ricardo O. **Rodríguez**

Luciana Micha Graduate in Political Science with a diploma of honor from the University of Buenos Aires, Mg in Neurolinguistic Programming and candidate for PhD in Political Science. She is currently director of the Center for International Policy Studies (CEPI) of the University of Buenos Aires. Lecturer at UBA and at the National Defence University (UNDEF) as Director and Lecturer of the Humanitarian Assistance Diploma Courses and Lecturer at the National Institute of Aeronautical and Space Law (INDAE) in the Diploma in International Law of Armed Conflict.

She has been a career official of the Ministry of Defense of Argentina since 2001, serving as National Director of Cooperation for Peace (2006-2010), Liaison with National Congress (2011-2012) Cultural Property Coordinator (2012-2015) and Coordinator of International Humanitarian Law (2015-2019). She currently serves at the University of National Defense, University Extension Secretariat.

Pablo Andres Farías is Commodore of the Argentine Air Force, in active retirement situation as Art 62. Military Aviator of the Attack Specialty, who served in the country, Argentine Antartic and abroad; performing functions in the areas of operations, education, flight instruction, operations, combat intelligence, strategic planning institutional/militar, projects and international relations (Argentina Embassy in Chile), among others. General Staff Officer (Air Forces of Argentina and Peru), with higher education (undergraduate, postgraduate, specializations and courses) in areas of interest related to National Defense, currently a candidate for a Doctor of Political Science. He developed the activity of General Staff, analysis, advisory and decision-making in Higher Air Force Agencies; the Joint Chiefs of Staff and the Ministry of Defense on strategic political, strategic-military, defense policy, strategic-military directives, Military capabilities and development of the National Defense White Card, among others.

María Vanina Martínez PhD in Computer Science (University of Maryland, USA) with a post-doctoral degree from Oxford University. She is coordinator of the Data Science and Artificial Intelligence Program at the Sadosky Foundation, researcher at the Institute of Computer Sciences (CONICET - UBA) in the area of Artificial Intelligence and professor of the Computer Department, Faculty of Exact and Natural Sciences, University of Buenos Aires, where she dictates the subject "Ethics & IA". She is also a member of the National Committee on Ethics in Science and Technology of the Ministry of Science, Technology and Productive Innovation and an adviser on Artificial Intelligence of the National Directorate for the Promotion of Scientific Policy (DNPPC), reporting to the Secretariat for Planning and Policies in Science, Technology and Innovation (SPYPCTEI). Shee is a member of the Stop Killer Robots campaign and the SEHLAC.

Ricardo Oscar Rodríguez PhD in Computer Science with a major in Artificial Intelligence. He is Associate Professor in the Department of Computer Science, FCEyN-UBA and member of the Institute of Computer Science (UBA-CONICET). His scientific works are part of the development of logical models for reasoning under incompleteness and uncertainty. He currently teaches "Ethics & AI" and is a member of SEHLAC and the Stop Killer Robots campaign. He has been co-chair & financial chair of IJCAI2015 in Buenos Aires from where the open letter of scientists against the use of AI in armament was launched.


APP
Asociación para Políticas Públicas


CEPI


SEHLAC
SEGURIDAD HUMANA
EN LATINOAMÉRICA Y EL CARIBE


CAMPAIGN TO STOP
KILLER ROBOTS