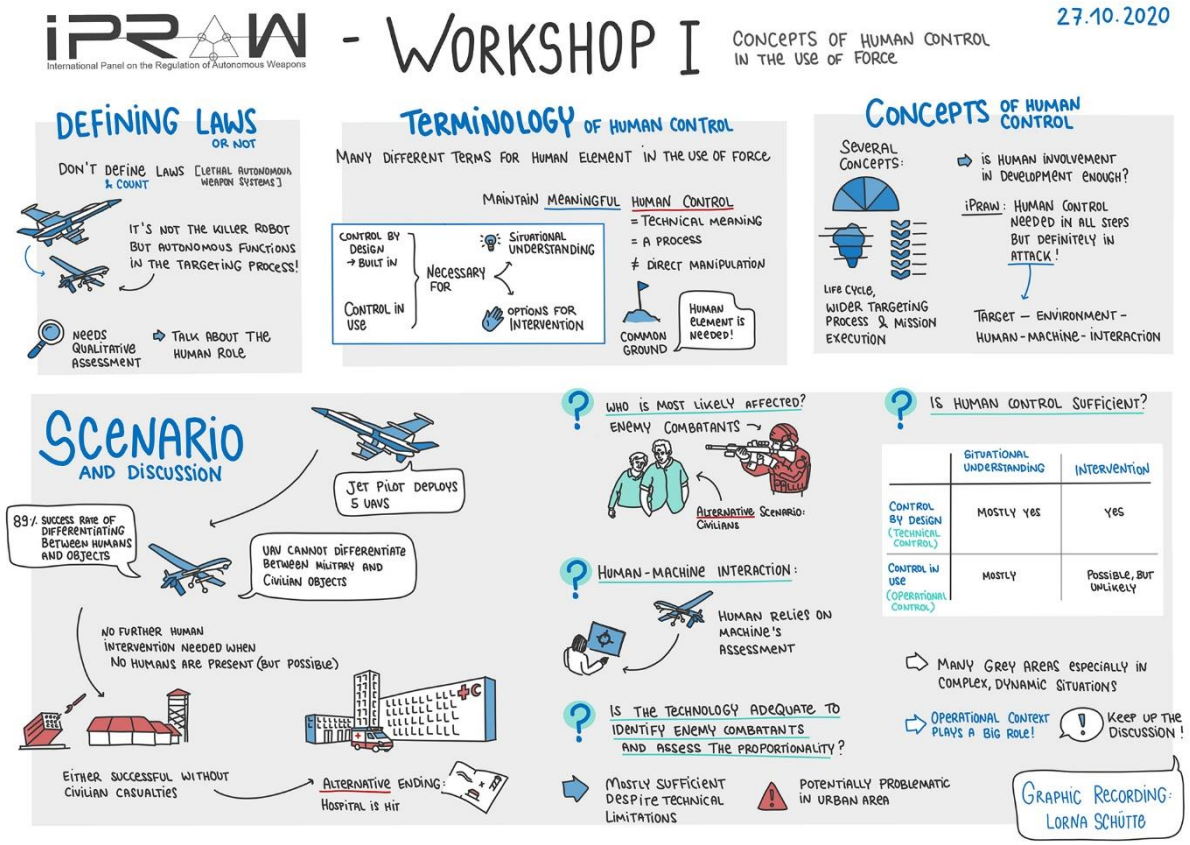
Hand Out for Workshop I – Concepts of Human Control

This paper summarizes the presentations given at the workshop "Concepts of Human Control in the Use of Force" by the International Panel on the Regulation of Autonomous Weapons on October 27, 2020.



DEFINITION OF LAWS – FOCUS ON HUMAN ELEMENT

So far, the CCW States Parties have not agreed on a common definition of LAWS for the GGE, but many states refer to the working definition by the International Committee of the Red Cross (ICRC). According to the ICRC, autonomous weapon systems denote

> "[a]ny weapon system with **autonomy** in its **critical functions**. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without **human intervention**."[1]

iPRAW recommends using the term lethal autonomous weapon systems as shorthand for various weapon platforms as well as systems of systems with machine 'autonomy' in the functions required to complete the targeting cycle. This stands in contrast to a categorical definition of LAWS. A categorical definition drawing on technical characteristics in an effort to separate LAWS from non-LAWS is unable to account for the already existing plethora of systems with autonomous/-automated functions and could, as technology progresses further, never be future-proof because almost every conceivable future weapon system can optionally be endowed with various autonomous functions.[2]

It is widely accepted that weapons can be deployed lawfully only if humans retain a certain degree of involvement in targeting decisions. What remains largely unclear, however, is the necessary type and degree of the human involvement, which will be discussed below.

TERMS TO DESCRIBE THE HUMAN ELEMENT – HUMAN CONTROL

There are different terms aiming to describe the human element in the context of LAWS. The GGE's **Guiding Principles** of 2019 refer to the human element by using terms such as "human responsibility" (Guiding Principle b), "human-machine interaction" (Guiding Principle c) as well as "accountability" (Guiding Principle d).[3] **Human-machine interaction** is a broad term referring to the question of the role humans play in the process of designing, developing, acquiring, deploying or using LAWS.

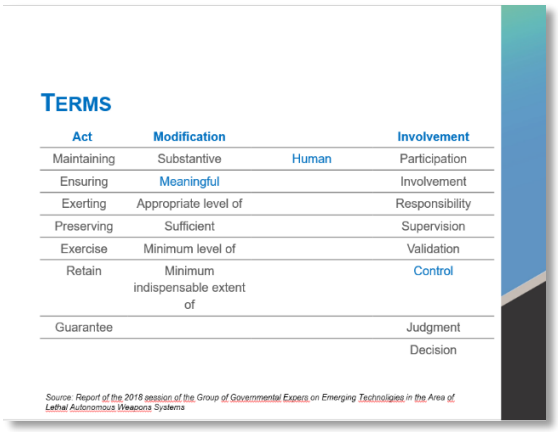
One option to describe the necessary type of interaction is the notion of the "**appropriate level of human judgment**". This concept does not require manual human manipulation of the weapon system "but rather broader human involvement in decisions about how, when, where, and why the weapon will be employed. This includes a human determination that the weapon will be used 'with appropriate care and in accordance with the law of war, applicable treaties, weapon system safety rules, and applicable rules of engagement.'"[4]

The NGO *Article 36* has coined the term **meaningful human control** to define an adequate form of human-machine interaction in the use of force. They divide the 'use of force' into three layers: (1) design, development, acquisition and training, (2) attack (as used in IHL and following a phase of operational or strategic planning), and (3) command structures and accountability. Human control has to be exerted on the lowest possible level during an attack, because "humans are the agents that a party to a conflict relies upon to engage in hostilities, and are the addressees of the [international humanitarian] law as written".5 Numerous CCW State Parties employed this term in their working papers and statements.

1 ICRC 2016.
2 iPRAW 2020.
3 UN Office at Geneva 2019, Annex IV.
4 USA - Department of Defense 2012, 2017.
5 See Roff and Moyes 2016.

Based on operational, legal, and ethical considerations, in iPRAW's understanding human control in the use of force requires at least situational understanding by the human operator/commander and the option to intervene built-in by design and available any time during use. The concept covers the whole **life cycle** of a weapon system from development specifications to mission planning to the attack and evaluation.[6] In iPRAW's understanding **human control does not necessarily equal direct manipulation**. The directness of the means whereby the agent seeks to control some object is related only contingently to the degree of control. Under some circumstances, manipulation that is more direct enables greater control, while in other circumstances the presence of an intervening mechanism might be the better option to reach the desired outcome. The increasing number of assisting systems does not necessarily increase precision, though, as they make the weapon system also more complex and possibly less predictable. A prudent balance of operational needs and situational understanding is crucial.

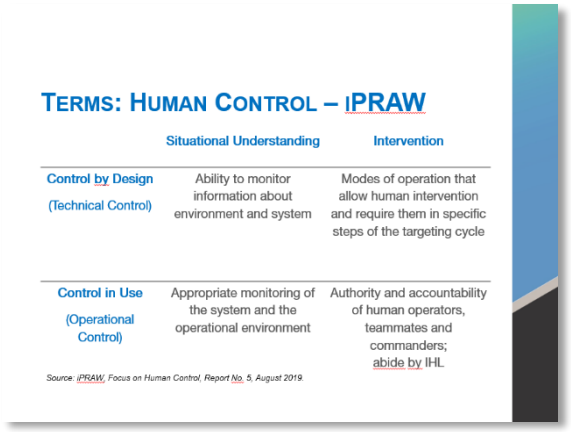
Although the various terms mentioned above take a different approach as regards the role humans play in the context of LAWS, it is at least their common denominator that a certain degree of **human involvement** in the decision making process is necessary to translate human intentions, judgments and legal assessments into operations. Even though the exact term might not be crucial to develop an effective regulation, the term human control appears to be quite helpful: it has a technical meaning in engineering, it describes a process rather than a singular event, and it does not necessarily equal direct manipulation.



TERMS

Act	Modification		Involvement
Maintaining	Substantive	Human	Participation
Ensuring	Meaningful		Involvement
Exerting	Appropriate level of		Responsibility
Preserving	Sufficient		Supervision
Exercise	Minimum level of		Validation
Retain	Minimum indispensable extent of		Control
Guarantee			Judgment
			Decision

Source: Report of the 2018 session of the Group of Governmental Expers on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems



TERMS: HUMAN CONTROL – IPRAW

	Situational Understanding	Intervention
Control by Design (Technical Control)	Ability to monitor information about environment and system	Modes of operation that allow human intervention and require them in specific steps of the targeting cycle
Control in Use (Operational Control)	Appropriate monitoring of the system and the operational environment	Authority and accountability of human operators, teammates and commanders; abide by IHL

Source: iPRAW, Focus on Human Control, Report No. 5, August 2019.

6 For details see iPRAW 2018.

Concepts to Operationalize Human Control

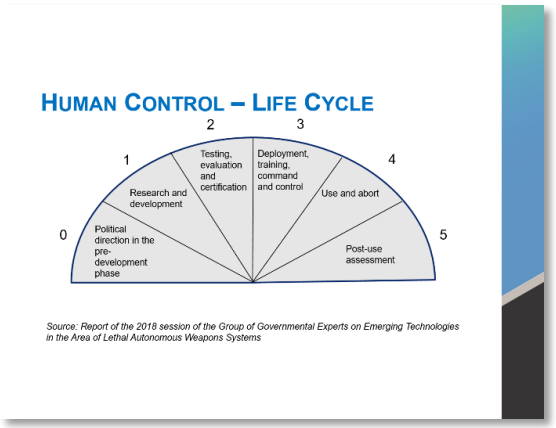
Different approaches aim at describing how human control can be implemented in the use of force. For the most part, they are not contradictory but build on each other and are quite compatible. This chapter gives an overview of a few of these concepts that have been especially influential for iPRAW's considerations.

Human Control across the Life Cycle of a Weapon System

In 2018, the GGE seized the idea of looking at the life cycle of a weapon as a whole when analyzing the notion of human control in the context of LAWS by presenting a 'sunrise diagram' visualizing the different human-machine touchpoints along the life cycle of LAWS. According to this concept, human control can be exerted in different forms and at different times in the various phases along the life cycle of a weapon. Proponents of this concept argue that the distribution of human control along the life cycle does not lead to a dilution of control because every step in the life cycle is intertwined with the notion of accountability.[7]

iPRAW agrees that many human decisions regarding the use of force are made beforehand ranging from development specification, the selection of training data, weapon reviews to mission planning and the attack. Nevertheless, iPRAW takes the view that human control cannot sufficiently be exerted at earlier stages in the life cycle of a weapon alone. Although human involvement along a weapon's lifecycle can inform and predefine decisions relating to a weapon's critical functions, human control still has to be retained during attack.[8]

The various concepts addressing human control differ regarding the question of when, how and by whom human control should be exerted. What they *do* have in common though is that they take into account the fact that LAWS are highly complex and that the broader decision-making process regarding their design, development, acquisition, deployment and use play a role to a certain extent. What remains rather contentious is the question of whether human control should be maintained during attack allowing humans to intervene if necessary.



7 See UN Office at Geneva 2018.
8 See iPRAW 2019.

Human Control in the Wider Targeting Process

UNIDIR stresses that critical decisions regarding the use of force are taken at various levels within the targeting cycle and argues that in order to consider human control holistically, one has to analyze aspects like legal reviews as well. They highlight the several stages of decision-making that lead to an attack ranging from the political level to mission execution at the tactical level.[9] This broader view has also been described as the wider targeting loop according to which target selection and engagement do not require options for human intervention if sufficient safeguards have been taken in advance.[10]

iPRAW also encourages states to take a broader view beyond the final stages of selecting and engaging targets. Decisions made at the various steps within the targeting cycle but also specifications in technology may inform or predefine decisions at the targeting level regarding target selection and engagement (see 'life cycle'). Nevertheless, iPRAW emphasizes that the most crucial factor is to maintain human control during attack even though decisions and technological arrangements made beforehand may impact human control during attack.

Human Control during Mission Execution / Attack

iPRAW as well as SIPRI/ICRC contend that human control has to be retained *during* attack. However, the question of when an attack starts and ends becomes difficult to define when autonomous targeting functions are involved. iPRAW argues that an attack commences when the final decision to use lethal force has been made. **In weapon systems with autonomous targeting functions, this decision does not include the selection of the target in a specific situation but a category of targets in a potentially wide geographical range over a longer period of time.** Therefore, the design of a system must allow human commanders to monitor information regarding the environment and the system itself (situational understanding) and to allow for intervention during use if necessary.

According to SIPRI/ICRC, three interdependent types of control measures have to be taken into account in order to guarantee human control in the respective operational context: controls on the weapon system's parameters of **use**, controls on the **environment** and controls through human-machine **interaction**. The design of a weapon – especially if using AI software or machine learning capabilities – can make a weapon system difficult to understand by the human operator and/or less predictable, limiting the control via the human-machine interaction.[11]

In line with these concepts, iPRAW takes the view that although it is possible to develop abstract minimum requirements for human control, the appropriate level or implementation of human control highly depends on the **operational context** in which LAWS are used. A "one-size-of-control-fits-all" seems hardly achievable.[12]

The International Panel on the Regulation of Autonomous Weapons (iPRAW) is coordinated by:
Stiftung Wissenschaft und Politik (SWP) – German Institute for International and Security Affairs
Ludwigkirchplatz 3-4, 10719 Berlin, Germany
This project is financially supported by the German Federal Foreign Office.
Find all reports and more information online at **www.ipraw.org**.

9 See Ekelhof and Persi Paoli 2020. For more details see Ekelhof 2019.
10 See Netherlands 2017. For another approach that considers the wider environment of military decisions see Verdiesen et al. 2020.
11 See Boulanin et al. 2020.
12 See iPRAW 2019, p. 6. Another helpful approach to look at the influence of the operational context is described by Amoroso et al. 2018 with five levels of human-machine interaction.

LITERATURE

Amoroso, Daniele; Sauer, Frank; Sharkey, Noel; Suchman, Lucy; Tamburrini, Guglielmo (2018): Autonomy in Weapon Systems. The Military Application of Artificial Intelligence as a Litmus Test for Germany's New Foreign and Security Policy. Heinrich Böll Foundation. Berlin. Available online at https://www.boell.de/sites/default/files/boell_autonomy-in-weapon-systems_v04_kommentierbar_1.pdf, checked on 7/28/2019.

Boulanin, Vincent; Davison, Neil; Goussac, Netta; Carlsson, Moa Peldán (2020): Limits on Autonomy in Weapon Systems. Identifying Practical Elements of Human Control. Available online at https://www.sipri.org/publications/2020/other-publications/limits-autonomy-weapon-systems-identifying-practical-elements-human-control-0, checked on 6/3/2020.

Ekelhof, Merel (2019): The Distributed Conduct of War. Reframing Debates on Autonomous Weapons, Human Control and Legal Compliance in Targeting. Amsterdam.

Ekelhof, Merel; Persi Paoli, Giacomo (2020): The Human Element in Decisions about the Use of Force. UNIDIR. Available online at https://unidir.org/sites/default/files/2020-03/UNIDIR_Iceberg_SinglePages_web.pdf, checked on 4/30/2020.

ICRC (2016): Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapon Systems. Available online at https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system, checked on 7/22/2019.

iPRAW (2018): Focus on the Human Machine Relation in LAWS. Stiftung Wissenschaft und Politik (SWP). Berlin. Available online at https://www.ipraw.org/wp-content/uploads/2018/03/2018-03-29_iPRAW_Focus-On-Report-3.pdf, checked on 8/6/2019.

iPRAW (2019): Focus on Human Control. Stiftung Wissenschaft und Politik (SWP). Berlin. Available online at https://www.ipraw.org/wp-content/uploads/2019/08/2019-08-09_iPRAW_HumanControl.pdf, checked on 9/16/2019.

iPRAW (2020): Commentary on the Guiding Principles. Stiftung Wissenschaft und Politik (SWP). Berlin. Available online at https://www.ipraw.org/wp-content/uploads/2020/09/iPRAW_Commentary_GuidingPrinciples.pdf, checked on 10/28/2020.

Netherlands (2017): Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. Examination of various dimensions of emerging technologies in the area of lethal autonomous weapons systems, in the context of the objectives and purposes of theConvention. CCW/GGE.1/2017/WP.2. Available online at https://undocs.org/ccw/gge.1/2017/WP.2, checked on 10/28/2020.

Roff, Heather; Moyes, Richard (2016): Meaningful Human Control, Artificial Intelligence and Autonomous Weapons. Briefing paper for delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS). Article 36. Available online at http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf, checked on 7/28/2019.

UN Office at Geneva (2018): Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 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. Available online at https://undocs.org/en/CCW/GGE.1/2018/3, checked on 10/28/2020.

UN Office at Geneva (2019): Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 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 online at https://undocs.org/en/CCW/GGE.1/2019/3, checked on 10/28/2020.

USA - Department of Defense (2012, 2017): Directive 3000.09. Available online at https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf, checked on 7/28/2019.

Verdiesen, Ilse; Sio, Filippo Santoni de; Dignum, Virginia (2020): Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight. In *Minds & Machines*, pp. 1–27.