Meaningful Human Control as an Exceptional Concept

Added Value and Common Ground for CCW Discussions on Lethal Autonomous Weapons Systems

Yeti Kakko

Yeti.Kakko@saferglobe.fi

January, 2022

EU Non-Proliferation and Disarmament Consortium

Promoting the European network of independent non-proliferation and disarmament think tanks

Disclaimer: This report has been prepared as part of a research internship at the SaferGlobe, funded by the European Union (EU) Non-Proliferation and Disarmament Consortium as part of a larger EU educational initiative aimed at building capacity in the next generation of scholars and practitioners in non-proliferation policy and programming. The views expressed in this paper are those of the author and do not necessarily reflect those of SaferGlobe, the EU Non-Proliferation and Disarmament Consortium or other members of the network.

Abstract:

The treaty negotiations in the Convention on Certain Conventional Weapons (CCW) on emerging technologies in the area of Lethal Autonomous Weapons Systems (LAWS) have been stalled by numerous terminological debates. A key debated concept is "meaningful human control." Although retaining human control is widely considered an essential factor in all autonomous weapons systems, no consensus has been found on whether meaningful human control brings additional value to regulation negotiations. At worst, this terminological deadlock can stall negotiations especially if contracting parties find other parties more focused on the actual regulation instead of terminological wrestling.

This paper presents "meaningful human control" as an exceptional concept where exceptionality means qualitative added value to understand aspects of LAWS. Exceptionality is elaborated through four key themes found in the extensive literature on LAWS. First, pressing a button as meaningful human control does not suffice since just any sort of control does not satisfy as meaningful control. The second theme is black-box systems and meaningful human control due to the opaqueness and unexplainability of advanced AI systems with multi-layered algorithms. The third key theme relates to high-risk systems, which require additional meaningful human control due to worse outcomes. High-risk AI systems also link to existing norms in the civilian AI domain, whereas low-risk systems require meaningful human control. The last thematic chapter revolves around the last resort argument. Although not in the focus of LAWS literature, the argument has lately found its way in regulatory texts relating to AI development. It requires attention due to its implications to what kind of use is allowed without any human control.

By recognizing the ways meaningful human control differs from just ordinary human control, the paper argues that the concept indeed provides additional value for the development and use of LAWS. It also shows that LAWS and meaningful human control are not mutually exclusive terms but that meaningful human control might even enable LAWS use. As States Parties to the CCW are stuck in terminological differences, the paper provides an opportunity to move beyond the deadlock and focus on regulative approaches instead of political poetry.

Keywords: Lethal Autonomous Weapons Systems, Artificial Intelligence, CCW, emerging technologies, black box, meaningful human control, disarmament

Table of Contents

1. Introduction

The negotiations for regulation on Lethal Autonomous Weapons Systems (LAWS) have not seen significant progress during the last four years. For many advocates of a ban treaty, the 6th Review Conference of the Convention on Certain Conventional Weapons (CCW) in December of 2021 was the last opportunity for the States Parties to the Treaty to take concrete steps towards regulation. The discussions have stalled due to different approaches between States Parties on the key concepts on LAWS. Many proponents for a pre-emptive ban have argued for a tactical change of venue to reach binding regulation. However, it demands a lot of resources and might mean the lack of interest in the process by the major military powers developing LAWS.[1] A treaty without the key actors is still a treaty and has some influence and can create customary international law. Still, it lacks the ability to globally regulate the development and use of a weapons category.

Despite calls to change of venue, many states have vowed to stay in CCW and see it as the best option for future discussions at least for now. However, development of consensus on the key terms of the debate and moving beyond the terminological poetry to focus on the regulation itself in the CCW is becoming all the more important.

One of the key terms in the debate has been *meaningful human control*, which was initially outlined by an A UK-based NGO Article36, in its 2014 report *Key Areas for Debate on Autonomous Weapons Systems*. The term has since been one of the most debated terms within CCW. While the States Parties to the negotiations on possible regulation have a widespread agreement on retaining human control over weapons systems, many see *human control as* insufficient to understand the requirements for the use of autonomous weapons.[2] Thus the concept of *meaningful* human control (MHC) has been used to fill the need for a deeper understanding of what it means to be in control.[3]

There remains little consensus on what exactly would consist as MHC and how it would be different conceptually from human control, which already has been at least latently agreed on – no country

[1] In *Crunch Time on Killer Robots* (2021), Human Rights Watch and International Human Rights Clinic call for a change of forum away from CCW if States are not able to proceed towards regulation on LAWS during the 6th RevCon of the CCW.

[2] Reaching Critical Will 2021(a); Human Rights Watch 2020.

[3] On meaningful human control, see, e.g., Bode & Watts 2021; Amoroso & Tamburrini 2021; Methani et. al. 2021; Santoni de Sio & Van den Hoven 2018; Wagner 2021; Horowitz & Scharre 2015.

has made public calls for weapons completely out of human control. Those calling for MHC argue that the existing IHL does not require enough human control for autonomous weapons systems, and the opponents to regulation argue vice-versa. I claim the concept of meaningful human control brings additional value and understanding to the regulatory discussions, especially if the concept is at least in some rudimentary form accepted instead of continuously challenged. The essay illustrates that despite the dissent amongst states parties to the CCW, the concept of *meaningful human control* is exceptional and thus a relevant part of the LAWS discussion.

From an extensive literature review on LAWS, four approaches emerge to make the case, through which "exceptional" is understood as additional meaning and value to the treaty discussions. Making a case for added value, the paper aims to push the LAWS discussion beyond the current terminological deadlock by pointing out in each approach how meaningful human control creates a better understanding of human control.

To expand on the argument about the usefulness of meaningful human control as a concept, I will discuss different ways to understand the concept in this essay. The themes emerging from the extensive literature on LAWS are 1) meaningfulness of pressing a button; 2) meaningful control of black-box systems; 3) requirement of meaningful control in high-risk situations; 4) meaningful control in the last resort use of LAWS. These sections work all as individual approaches but tie to each other by following the themes from the previous.

Due to the nature of the themes chosen, the case against exceptionalist military AI development is also made. By addressing civilian applications of AI, I argue for mutual control requirements with military and civilian AI applications. Considering what amount and kind of human control we require in civilian use of AI is noteworthy due to the unique role in the development of AI. With previous big leaps in technology, the military has generally been the inventor, developer, and first user of the given technology. Technologies have then proliferated to the civilian sphere. The private sector primarily drives AI development, and the armies either compete with them or buy innovations popping up in the civilian sphere.[4] The turnaround means the norms and regulation in the latter takes the driver's seat instead of being a passive passenger. Some worry regulating military domain application would hamper the civilian industry.[5] It is a two-way street, however, and regulation and

[4] See for example Altmann & Sauer 2017; Scharre 2018; Singer 2009.

[5] The worry of overregulation is apparent in the language used by many, see e.g., UNESCO 2021; European Parliament 2021; US DoD 2018; The White House 2019; BDI 2021.

requirement for meaningful human control in the other would most likely spill over to the other as well.

Before the thematic approaches, the next chapter briefly introduces LAWS as an emerging topic and the CCW as the forum for treaty negotiations. The chapter brings forth the issue of human control as a concept and meaningful human control as its own unique way to approach the question.

2. LAWS, the CCW, and Human Control

LAWS has emerged in the field of international law and international relations relatively recently. It gained broader attention only in the 2010s. In the United Nations, the first mention was made in a report by Special Rapporteur Christopher Heyns to the UNHRC in 2013. [6] Soon after, states, led by France, moved the discussions on autonomous weapons from the UNHRC to CCW, thus framing the topic as a security matter instead of a human rights issue. This move has been seen as surprising as, typically, human rights have been the driving force for disarmament treaties.[7] However, the push for regulation and the call for human control has been from the human rights perspective from the start of CCW discussions.[8]

The discussions in the CCW on LAWS started as informal but have been held as formal Group of Governmental Experts (GGE) meetings since 2017. CCW is an umbrella convention for multiple different weapons categories, and new protocols have been added to it under the initial three in 1980. The convention has 125 High Contracting Parties and four signatory states. Decisions are made by consensus, creating issues in matters that divide opinions. Successes, however, are proven to be possible under the CCW protocols, such as banning blinding lasers (Protocol II) or incendiary weapons (Protocol III).[9] The consensus has proved troublesome on LAWS as the States Parties seem not to find enough common ground to advance the discussions towards a legally binding treaty. The

[6] UN (2013), A/HRC/23/47.

[7] Barbe & Babell (2019), 136.

[8] ICRAC 2009; Alston 2011; Human Rights Watch 2012.

[9] CCW Protocols are as follows: I) Non-detectable fragments (1980); II) Mines, Booby Traps and Other Devices (1980); III) Incendiary weapons (1980); IV) Blinding Laser Weapons (1998); V) Unexploded remnants of War (2006).

most significant milestone so far was the 11 guiding principles, on which the states reached consensus in the Final Report of the CCW GGE in 2019.

Although consensus has not been reached in many essential concepts such as autonomy or what would consist as a weapons system under possible regulation throughout the treaty negotiations in the CCW, a broad agreement on retaining some control over the use of force has been evident. None of the States Parties to the convention have publicly endorsed either development or use of weapons systems entirely out of human control. This view is reflected in the 11 guiding principles on LAWS set out in the Final Report of the CCW GGE in 2019. The principles explicitly state human control in paragraphs:

> 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 complies 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."

and

> 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."[10]

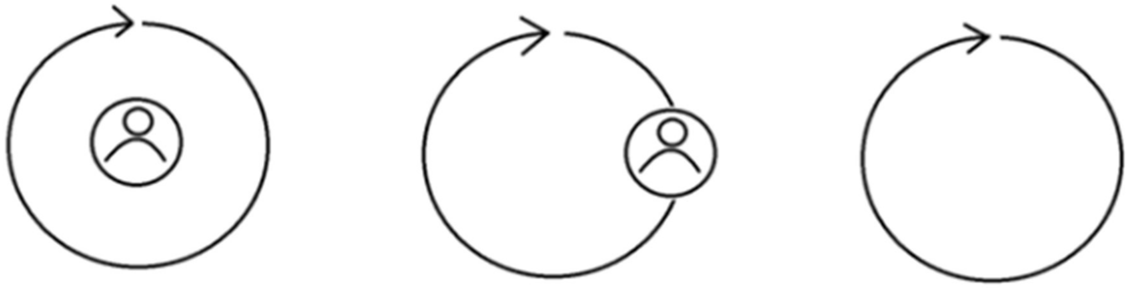
Not being in control is commonly referred to as humans' inability to affect the system's operations after it is activated, or the widely used phrase, out of the loop. The term "loop," commonly shortened term for OODA-loop, refers to a typology of a human being "in," "on," or "out" of the decision-making loop (Table 1.) [11]

[10] Final report of the CCW GGE on LAWS 2019, CCW/MSP/2019/9 Annex III.

[11] Observe, Orientate, Decide, Act (OODA) was created by US colonel John Boyd during the Korean war (McIntosh 2011).

Figure 1.

Human and the loop



Human in-the-loop	**Human on-the-loop**	**Human out-of-the-loop**
A human is needed for the processes.	The machine makes most of the work. Human functions as an overseer to make sure the system operates as intended. The overseer can interrupt and stop the process if needed.	Human is not needed. The system has become precise and autonomous enough to complete the process by itself.

Table 1. Human and the loop. (Illustration by the author).

The consensus ends with the understanding of the role of the human. Human control as a principle sounds good, but the states have differing views on the term for sufficient human control and what the terms would then mean for possible regulation.[12] Instead of meaningful human control, some have chosen to use other terminology such as "appropriate human involvement"[13], "meaningful

[12] For instance, Russia has claimed meaningful human control merely politicizes the discussion (CCW/GGE.1/2021/WP.1). The United States would like to see 'human control' replaced with "responsible governance" (US statement in the 4th meeting of the 3rd session of the CCW GGE 2021)

[13] Israel's statement during the 1st meeting of the 3rd session in the CCW GGE.

human involvement"[14], or "human-machine-interaction,"[15]. Still, the understanding of these concepts remains essentially the same as all these concepts refer to the role of the human.[16] The apparent polyphony is demonstrated in the figure provided in the 2018 Final report of the CCW GGE on the used terms during the discussions (Table 2.). Although the terminology has seen some variations after 2018, the table illustrates the vast differences in the chosen language.

Maintaining	Substantive	**Human**	Participation
Ensuring	**Meaningful**		Involvement
Exerting	Appropriate		Responsibility
Preserving	Sufficient		Supervision
	Minimum level of		Validation
	Minimum indispensable extent of		**Control**
			Judgment
			Decision

Table 2. The terminology used on human control in the 2018 CCW GGE discussions on LAWS, according to the Final Document submitted by the Chair of CCW GGE on LAWS 2018.

Some states, such as the United States, have argued in their statements to the CCW that using a system with humans entirely out of the loop is not strategically sound. The results from using such a

[14] UK's statement during the 1st meeting of the 3rd meeting in the CCW GGE.

[15] Final report of the CCW GGE, CCW/GGE.1/2019/3.

[16] Curiously, the briefing paper by Article36 states that "The CCW meeting of experts offers an important opportunity for government delegations to: Reaffirm that meaningful *(or sufficient of appropriate)* [emphasis added] human control must be exercised over the use of weapons (…)". Thus, making a difference between the terms in the original context makes little to no sense.

system would be unexpected and, thus, indiscriminate in its effects. These sorts of systems are already prohibited by international humanitarian law (IHL).[17] The IHL already categorically prohibits intrinsically indiscriminate weapons, and as such, should be a clear cut in the human involvement in the systems. However, it is unclear at which point of the chain of command and use of the system the human control should be applied. Here the ways depart for many of the members of the LAWS discussions. For example, opponents of a legally binding treaty argue that no additional regulation is needed as the mentioned principle in the IHL is sufficient in itself.[18] In contrast, some see sufficient human control as someone pressing a button before initiating attacks and some as more in-depth, continuous principle throughout the chain of command and control of military operations.[19] This essay shows that merely pressing a button is not enough, and as such, approaches such as the latter are more useful approaches to understand control over LAWS.

Here the concept of meaningful human control provides the next steps. As there are differing views on what constitutes human control, the thematic approaches demonstrate how and why meaningful human control is of exceptional value to LAWS regulation, development, and use by approaching the issue through key themes emergent in the LAWS literature. Using a single term, meaningful human control, we set out a common understanding for the concept and need for human control and move beyond the dispute over semantics. Although MHC and LAWS may seem mutually exclusive, they do not have to be. By focusing on the exceptional and additional value MHC provides, the conclusion is it ensures the legality and morality of LAWS instead of creating a demand for total prohibition of both development and use of them.

3. Pressing A Button Is Not Necessarily Meaningful Human Control

As stated, some see the existing codes of conduct for the operation of weapons systems as sufficient for future weapons systems. It is also noteworthy that current *legal* weapons do not operate outside human control, and a human is in the loop to some extent in all situations. Why would we then need additional requirements for human control for autonomous weapons systems? Meaningful human control embodies the idea that regardless of the automation of any given weapons system, a human

[17] ICRC (2021), IHL rule 71. Weapons That are by Nature indiscriminate.

[18] Future of Life 2021; Ford 2017; US statement in CCW GGE 28.8.2018.

[19] See e.g., UNIDIR 2020.

must be *meaningfully* in control of the system. Pressing a button, for example, does meet the requirements for human control but not the requirement of *meaningful* human control.

For the sake of illustration, let's consider a box with a screen and a system trained to give suggestions for action. The role of a human is to either press the yes or no button when a suggestion pops up on the screen. Here a human decides but has little aside the limited options given by the system. Without providing a reason to doubt the machine's capability, the person in the box would agree with the suggestions. There would be humans in control of the decisions in the example provided, but it is not exactly meaningful, as argued above.

Automation bias is especially relevant since humans tend to trust machines and make the decisions they propose. [20] Good faith may lead to decisions where we do not intervene in the system's actions even when we probably should and had the possibility. Trusting in machines is not a new phenomenon [21] and it makes sense, for instance, while using voice-controlled assistance such as Siri or Alexa, grammar correction software, or translation applications. These examples are straightforward systems working in simple environments. Self-driving cars and airplane autopilots are examples of systems working in more complex environments. Still, there remains a human in the system, and there is always a way to overrule the system's decisions.

Automation bias and false positives have on various occasions proven to cause unfortunate outcomes. A commonly referred case is the US military's shooting down of Iranian civilian aircraft in 1988 due to a recognition error in its AEGIS air defense system, which misidentified the Iran Air Flight 655 as an enemy F-14.[22] Cases of misfires and fratricides illustrate the need for more human control than just pressing a yes or no.

The practice in multiple military systems already follows the human and a button protocol—missile defense and radar systems work as described in the box and a button example.[23] The argument still stands; it is human control but not meaningful because machines tend to be brittle and erroneous, and

[20] Cummings 2004; UNIDIR 2018.

[21] For example, in his book, Singer gives a historical overview of the development of robotics and illustrates the fact we tend to believe robots work as we think we do, although they do not. (Singer 2009, 42-65).

[22] Scharre (2018), 169–170.

[23] Bode & Watts (2021), 26-28.

we as humans tend to believe the machines. Existing practices of human control in weapons systems have not been an issue for LAWS ban treaty proponents. This view is reflected mainly in the statements by the United States in the CCW.[24] However, it is stated that some existing systems might be in the scope of regulation due to the control question. For instance, existing air and missile defense systems include autonomous functions, thus leading to question of whether they should also be strictly regulated under a LAWS treaty. [25] [26]

Thus, another question arises; should autonomous weapons systems have a higher bar of requirements than existing human-controlled systems? This question is frequently brought up in CCW discussions and other LAWS forums. Humans get tired, have personal biases, and generally are not perfect at anything humans do. Humans don't have a clean track record in the history of war either. Does the control need to be meaningful if the machine/system makes accurate estimates and suggestions without more human control?

If a human in the box makes the right choice 99 times out of 100 and an AI system achieves the same result, it would be difficult to argue against machines doing the same task instead of humans based on the accuracy of decision making. However, the accountability for any accidents resulting from that 1% error marginal cannot be transferred to a machine, whereas a human is responsible for all decisions made, even erroneous ones. [27] A human pushing the button is not needed for the sake of securing sufficient results but to ensure a human is in or on the loop and thus legally and ethically responsible for the outcomes. Meaningful human control would implicate the need for sufficient human control to hold a person morally and legally liable when using autonomous weapons systems – not just being part of the loop. As AI-enabled systems are developed to be increasingly complex, they also may become opaque and unexplainable for the user, thus further demanding meaningful control.

[24] For example, see US statements to the CCW 10.4.2018; 26.3.2019.

[25] Bode & Watts 2021: Campaign to Stop Killer Robots 2020.

[26] Systems such as South Korea's sentry gun SGR-A1, Israel's Iron Dome and Harpy-drones, and Turkey's Kargu-drones already include many autonomous functions.

[27] Majority of States, NGOs, and academics, regardless of their position towards LAWS and a ban treaty, have agreed that the moral or legal accountability or responsibility cannot be transferred to a machine. More on accountability see e.g., ICRC 2021; Crootof 2016; Marquelis 2016; Simpson & Müller 2016; Roff 2014.

4. Autonomy Is Not Just a "Glorified Excel"

Existing weapons systems operate with a reasonably understandable logic, such as pulling a trigger result in the weapon being fired. While some claim that AI is a qualitative leap towards a different operating logic, most share a contrary view that despite the advances in AI, decision-making is still based this the same logic. The decision made by machines can, at least theoretically, be divided into simple "if A, then B" solutions. Even the most complex deep learning systems currently work with deductive reasoning processes, *modus ponens*, with millions of layers of nuanced "if A, then B".[28] In other words, at their essence, these systems are akin to glorified excels with the added ability to crunch numbers. Here two essential avenues open for meaningful human control and machine decision making. First, if the systems follow deductive logical reasoning in their decision-making, in theory, the decision-making trees are explorable and traceable back to origins. The 'glorified excel' understanding diminishes the role of the human as the calculations, like in Excel, could be fully automated with minimal human input. However, a more applicable way to understand autonomous systems is the 'black box' approach which problematizes the transparency of the decision-making tree due to the millions of layers of unknown decisions. This chapter will first elaborate on the concept of autonomous systems as little more than glorified excels. Then it will focus on the black box approach and its advantages of use as a basis of further analysis.

Autonomous systems as "glorified Excels"

For some, autonomous systems are little more than "glorified excels" in other words advanced calculations using enormous amounts of data just like a very extensive Excel. There is no explicit requirement for human control in using Excel sheets, although accountants and data analysts work Excels. Some expect accountants to be the first group of people to lose their jobs to AI.[29] Though only speculation, the idea of losing our jobs to AI is not far-fetched. Algorithmic decision-making is just calculations based on inputs. Thus, claiming a different need for human control in Excel sheets and in more advanced systems also claims a significant qualitative leap between 'normal' number

[28] See e.g., Larson (2021), *The Myth of Artificial Intelligence – Why Computers Can't Think the Way We Do.*

[29] This was brought up by professor Teemu Roos in a short interview between him and I on artificial intelligence and human control. The origin of this claim most likely originates from news of American media organization NPR's calculator, that predicted accountants have around 95% chance to lose their job to AI.

crunching and AI number crunching. Although tempting, the claim of a fundamental difference between the two easily slips into a dystopian singularity argument, where AI systems are mystified and often seen as omnipotent as in sci-fi such as Terminator, 2001: Space Odyssey, or the Portal videogame series.

The consideration of possible development trajectories for autonomous weapons systems based on assumptions based on real-life cases with a touch of science fiction is a significant reason for the seemingly large gap between proponents and opponents of the LAWS ban treaty. For example, the discussions in CCW GGE between 2017-2019 saw the dichotomy between the USA and treaty advocates, since the US cleverly called for "real-life examples" of autonomous systems as the basis of the discussion and then continued to be the one giving those real-life examples, as no-one else had the technical capability to do that.[30] Distinguishing between reality and speculation was a useful rhetorical tool to dismiss any additional meaningful human control claims. The divide may be bridged to some extent through enhancing understanding of AI and focusing on practical, current, considerations. Autonomy or autonomous systems would not be seen as otherwise qualitatively different. The "keep control" argumentation would then most likely decrease as well or disappear altogether since it would be strange to require additional meaningful human control in systems that are not any different from existing ones.

Devaluing the need for human control, however, would lead the negotiations in a never-ending stalemate between those who believe IHL is sufficient for LAWS regulation and those who do not. It would also reduce the input of ethicists and other relevant participants and focus the debate solely on legal questions. Claiming that only the legal aspect matters would implicate a claim that the advocates of meaningful human control merely use it as an advocative tool to hinder the discussion and steer it towards something they favor and can take part in. In addition, advanced AI systems require more in-depth understanding and demand for meaningful human control. A common characteristic of advanced AI systems is opaqueness. The opaqueness is commonly referred to as the black box effect, which would then require additional and more meaningful human control.

Autonomous systems and the "Black box"

[30] UN 2017, CCW/GGE.1/2017/WP.6; UN 2018, CCW/GGE.2/2018/WP.4; US statement in the CCW GGE 27.3.2019; US statement in the CCW GGE 9.4.2018.

When people refer to AI decision-making as being qualitatively different, an issue commonly referred to as the ‘black box’ comes up. The term is used for the multilayered decision-making tree, from which humans can see the outcomes but from which it is difficult to precisely explain or understand how the machine ends up in that exact outcome.[31] Here *meaningful human control* brings forth the lack of understanding and explainability of an autonomous system and its stacks of unexplainable layers. The consideration of the "black box” is between the engineering-focused understanding of AI only as calculations and the dystopian understandings of AI having the potential of e.g., achieving consciousness and surpassing human intelligence.

Black box systems are, in the end, calculations and in their very core the same “if A, then B” reasonings. However, the millions of layers of these calculations create opaque systems. With increasing opaqueness, the awareness of the logical reasoning behind outcomes decreases. It is somewhat like the question of human free will; if it would be possible to know and identify all the factors that affect human decision-making, in theory, it would also be possible to know what is going to happen in the future. This deterministic model has been challenged by many, and the idea of not having free will may be disturbing.[32] The same pattern applies to black-box AI systems as well. In the case of AI, there are reasons for the outcomes, but no tools to explain them transparently, thus making it difficult to understand the machine’s decision-making. It has been evident in many practical uses of such systems, such as AlphaGo with its unforeseen moves in the game Go, leading to its victory over the human champion.[33] Also, many picture recognition systems end up in their results in quite unexplainable ways.[34]

For these kinds of machines to be meaningfully human-controlled, one precondition could be the technical skill and understanding for the user to operate said systems. As with the “pressing a button” argument, although a human operator would be in or on the loop, without understanding the machine other than “I’ve read that it works,” would not be concise as meaningful human control since the person pressing the button could not explain why something went wrong if it did. Without proficiency in advanced AI and the system in question, the person operating it could also not anticipate the

[31] Holland (2020) makes the distinction between ”interpretable” or “transparent” models and “opaque” systems, which are not inherently understandable.

[32] More on free will, see Dilman (1999).

[33] DeepMind (2021).

[34] Scharre (2018), 180–188.

system's behavior in surprising situations, thus making it difficult to abort the mission at the earliest moment possible.

To make the issue even more complex, there is a difference between *understanding* and *explaining* the machine. In UNIDIR's recent publication on the topic of Black Box, the author Arthur Holland uses a toaster analogy: "*most people probably do not know exactly how their toaster works, but they do have a robust mental model of how it turns bread into toast*".[35] The analogy vividly explains the difference between the two. Here the call for meaningful human control is not limited to only the end-user of the system but covers the entire command chain for the use of the system. Who should be technically savvy enough to be authorized to use these systems? The soldier during the operation, the commanding officer, the military leadership, or the political leaders allowing the use of force with these systems?

Even with limited understanding of the opaque mechanisms behind meaningful human control could be achieved through a self-explaining AI. For instance, an additional narrow AI to create explanations for the more complex AI. If this is achieved, a deeper look into the combination would be required to ensure it suffices as meaningful *human* control, as "who watches the watchmen" is not a new phenomenon. However, MHC as a concept would indeed be exceptional and bring additional value to the conversation as clarifying it would create qualitative prerequisites for the development and use of LAWS including some black-box AI to be explainable and understandable. To ensure sufficient control over LAWS, *meaningful* human control is needed, not just human control. Meaningfulness is, even more, required the more risks are involved in the use of such systems.

5. High Risk – More Meaningful Human Control

From the perspective of meaningful human control, the issue with unexplainable and opaque AI systems also relates to the risk involved in using those systems. The question of risk has also been one of the main arguments against the development of LAWS as it might lower the threshold of war [36] or the machines may "go rogue". [37] AI may also be considered too unreliable to be delegated tasks,

[35] Holland (2020), 10.

[36] Kovic 2018; Reaching Critical Will 2021(b).

[37] On the threshold of war and riskless war, see Future of Life 2021; Henriksen & Ringsmose 2015; Human Rights Watch 2015. Kahn 2002.

especially on the ever-changing battlefield.[38] The last is high relevance for meaningful human control; the moral and legal liability cannot be transferred to humans. Additionally, the systems are unexplainable even in controlled and straightforward environments, let alone in dynamic environments where the objectives are not straightforward, and multitasking is required from the system. To illustrate issues arising even in a simple AI environment, a recent case example from Finland is analyzed.

Finland uses automated decision-making in many instances, such as study grants and taxes. In 2018 a case against the national tax administration was brought up due to an automated system. Annually around 120 000 tax decisions are done by the system without a single human being involved at any point (other than the private person filing their taxes online). It resulted in situations where persons were obliged to pay their taxes twice or were wrongly sent due to notifications and threatened with tax increases. The Deputy Ombudsman went through the cases and ruled the automated procedure of the Finnish Tax Administration was contrary to the Finnish Constitution.

The decision was based on two main factors; first and foremost, the Finnish law requires a human in the loop in the sense that the name and contact information of that human need to be available for each decision. In the automated decision, the contact number was the general service number. The only contact official available was not affiliated with decision-making and thus was not aware of any details. The Deputy Ombudsman ruled it to be against the principles of good public governance to make taxing decisions only based on statistics. The decision also elaborated the algorithms used in the system must be transparent enough for the person using the system (be it the official or the person filing their taxes) to understand it. In addition, the person affected by an unlawful act must have the right to seek recompense from the public authority.[39]

The example provides one basis for assuming, that civilian legal frameworks and thus socially accepted norms already require a certain amount of human control in the computer. Further, there is an implication that of a robust view of what machines are and are not allowed to do without human control. If humans are unwilling to let automated systems make decisions without human control in the civilian domain, then a similar level of control could be expected to exist in the military environment.

[38] See, e.g., McDougall 2019.

[39] Decision by Finland's Deputy Ombudsman EOAK/3379/2018.

Asking for meaningful human control thus does bring additional value because a high threshold of human control for AI in all environments is required.

The military domain, however, makes the need for meaningful human control more apparent. The amount of risk involved in the outcomes sharply distinguishes most civilian and military applications apart. Unlike unfortunate taxing decisions, the decisions to use weapons systems do not only have unlawful but irrevocable consequences as well.[40] For LAWS, the emphasis is for L, as in Lethal. "Ask forgiveness, not permissions" is a poor rule of thumb for waging war with autonomous weapons systems.[41]

What increases the threshold for risk is also the complex environment AI applications would be used in. To begin with, only narrow AIs exist. Narrow AI systems are good in the few tasks they are given in a controlled environment, but they fail in other tasks and in other environments. Systems capable of multitasking, improvising, and working in complex environments are called General AI or Advanced General AI (AGI). Although some expect the eventual dawn of AGI to bring significant benefits, there are no known algorithms for AGI or a clear engineering route there.[42] A narrow AI system becomes brittle if thrown into a dynamic environment. If humans are not okay with AI making decisions on taxation, letting an AI make decisions on the lethal force in complex conflict environments seems strange. The risk for an unwanted outcome is high. The question would then be how could one ensure a tolerable error margin in these complex environments? And how would the risk threshold be assessed?

If these AI systems were to be used to their fullest extent, the risk tolerance at first would seem to have rather strict margins. States would likely need to legitimize their use and positive proof of LAWS decreases the risk of facing bad publicity. The margin for risk tolerance would be adjusted through rigorous research and development, as already claimed by proponents of LAWS in the

[40] Many states do not have the same social safety net, and thus even these kinds of automated systems pose actual risks for human wellbeing. For this reason, De Sio and Van den Hoven (2018) make the claim for meaningful human control to be understood in every domain where AI could affect any activity linked to basic human rights such as life, physical integrity or freedom and privacy.

[41] However, it seems this is the way forward for the regulation of autonomous weapons systems, as the big players are not willing to make steps towards any regulation. Thus, the result will most likely be a) no regulation at all, or b) regulation after the development and use, if the operational use proves to be immoral or illegal.

[42] Lee (2018), 142.

CCW GGE so far. As with acquiring any weapons system, the article 36 requirements stay.[43] If, however, lethal autonomous weapons systems operate soundly and without unlawful outcomes, the public acceptance of higher risk tolerances is likely to grow with increased trust in the systems.

Thus, calling for meaningful human control preemptively strengthens industrial standards and encourages companies to incorporate the principles in their product design. Enabling this incorporation is necessary especially in view of what some have called the "AI arms race"[44], referring to the very quick development of military AI between major military powers such as the United States, Russia, and China. The prospect of a global race to be the new world leader of state-of-the-art war technology makes it tempting to lower the risk threshold of LAWS. It is likely due to other contemporary technological developments in the military sphere, for instance, hypersonic missiles. Increasing speed in warfare is expected. Humans fall behind in the speed of warfare; humans simply cannot compete with AI or hypersonic weapons.

However, developing these systems to stay on par with others makes. Calling for meaningful human control during the regulatory phase of weapons development is necessary, if not almost mandatory. Exceptional situations require exceptional means. Military logic could easily dismiss the calls, but militaries would then face the negative PR for neglecting the obvious calls by the public.

6. Last resort as legal use but with meaningless control

Although in most cases, the use of LAWS can be expected to be relatively mundane, the discussion of the last resort highlights the need to nevertheless develop meaningful human control. Exceptional circumstances often, however, call for exceptional measures. If the situation is bleak, cannot then any means necessary be used? The understanding of the "last resort" is based on the just war theories but applies well also to LAWS, even though analysis is limited. Just war theory and LAWS have been analyzed in some instances [45], but the last resort argument has not been extensively studied in relation to LAWS. However, the concept is especially familiar from the discussion related to other

[43] Article 36 of the Protocol I to the Geneva Conventions: "*In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.*"

[44] Roff 2019; Scharre 2019.Geist 2016; Hoffman 2018.

[45] Simpson & Müller 2016.

weapons classes, where weapons may both be prohibited but also allowed for use in last resort-scenarios. As this understanding of qualified prohibition already exists, it would be applied to LAWS where AI-based killing would not be allowed at all or without meaningful human control, except in last resort cases.

The last resort argument, which refers to the means in conflict when other measures fail to bring results and self-defense creates further possibility of analyzing the use of LAWS in times of duress. Pooling last resort and LAWS creates implications on human control, which require attention for two reasons. First, their use as a last resort means they would be allowed to be developed in the first place as using something as a last resort requires the means to use it use in the first place as well. Second, should LAWS regulation be understood in the context of conventional weapons or be treated a specific category in its own right? Choosing either one of these approaches means different things on how we are to understand meaningful human control, as we put different emphasis on the use of different weapon categories.

Last Resort has not been extensively analyzed in LAWS literature, but it was mentioned in the European Parliament's resolution on artificial intelligence:

> "The use of lethal autonomous weapon systems raises fundamental ethical and legal questions about the ability of humans to control these systems. Such systems should meet a minimum set of requirements and *be used as a last resort* [emphasis added]. They should only be considered lawful if they are subject to strict human control" [46].

It is difficult to prove which situations make last resort measures just and are all measures just even in those situations. Allowing the use of LAWS as a last resort thus begs the question of the requirements of keeping control over these systems. An excellent example of questionable last resort measures is the "dead hand mechanic" developed by the Soviets, which is still currently in use in Russia. The "dead hand" assures mutual nuclear destruction in a situation where Russia would be under almost total nuclear annihilation and there would be no one to press the launch button, hence the "dead hand". These kinds of systems can be effective deterrence systems. Making others aware of the system also moves responsibility to others as their action will lead to the activation of the system. There is no human in this loop, and hence no meaningful control but simultaneously the system is only

[46] EU (2020).

to be used in the extremely exceptional circumstance of near annihilation. However, without a proper last-resort situation, an all-out uncontrolled nuclear launch would be inherently *mala in se*. A LAWS with the same function should then include an on-off switch between meaningfully controlled use and uncontrolled use. Developing a useful, properly working uncontrolled system, however, would imply developing one should be allowed in the first place. By default, it would mean, in addition, that meaningful human control would not be needed in all cases where LAWS is used.

Trying to articulate the boundaries of last resort situations at this time of LAWS regulation is difficult and thus excluded from the negotiations. The attempt to create a threshold for last resort used for LAWS would just add another dimension to the same conversation. Also, last resorts are always exceptional and thus open for interpretation on a case-by-case basis. As with any regulatory language, it should be concise instead of subject to interpretation. Using clauses such as a last resort, the European Parliament leaves gates open for loopholes. Loopholes, however, are dangerous as they leave the door open for those not interested in an international regulation regime. If the aim is to create binding regulation for the development and use of LAWS or a total prohibition of either, no loopholes should be left. Requirement for meaningful human control would null the possibility of last resort use and thus close the loophole left by last resort arguments.

Creating parallels between existing and new weapons systems could help the regulatory framework to use previously accepted language. Out of the existing weapons, however, so far, the discussions on LAWS have seen parallels drawn mostly between LAWS and landmines due to their innate nature of autonomously ‘choosing’ their targets. The similarities are also evident in the advocacy campaigning as they use the same methods in the LAWS campaign as was used in the discussions leading up to the drafting and signing of the Ottawa treaty back in 1997.[47] If, however, LAWS should be treated in their own specific category, the approach and the understanding of the need for human control will need to be reformulated.

Considerations related to other weapons classes and the last resort argument may further strengthen the understanding of LAWS, as military AI development has already been called the next revolution in military affairs after nuclear weapons. However, a last resort or deterrence arguments have not been used in the CCW GGE meetings to defend LAWS although there is the political will to do so, as proven by the Parliament resolution. Further developing the concept of meaningful human control in

[47] Cottrell 2009.

the context of the newly elevated LAWS brings forward another question. If the approval of the use of LAWS is that of nuclear weapons, the same requirements should apply for both. Thus, asking for meaningful human control – or 'strict human control' as worded in the resolution – for LAWS would imply the same kind of requirement forother weapons classes as well. If meaningful human control is solely required from LAWS but not other weapons classes, there is a danger of creating contradictory incentives for weapons development and the classification of AI-assisted weapons as other kinds of weapons. The concept of meaningful human control is thus an exceptional approach in the last resort argument.

8. Conclusions

The exploration of meaningful human control through the thematic approaches has reinforced the argument that meaningful human control (MHC) as a term brings additional value to the understanding of LAWS and its regulation. Albeit *meaningful* has different connotations and must be implemented on a case-by-case basis, the concept itself should be accepted. The exceptionality shines especially in the operational context of LAWS, as the contemporary practice of human control in weapons systems does not adequately satisfy the need for meaningful human control. Pressing a button requires additional measures to be taken for the control to be truly meaningful. Human control by itself may be regarded as sufficient for legal liability but does not meet the criteria for moral liability. MHC is also required for the satisfactory use of advanced AI systems, which are opaque, unexplainable, or have other black box characteristics. If LAWS instead are explainable and are not black boxes, MHC is not required. As military applications of autonomous AI inherently have a higher risk for irrevocable outcomes, is MHC essentially needed. The sentiment is further cemented by norms for AI use and human involvement in the civilian sphere, where human is required in the loop. Military applications of AI should not be allowed to have an exceptional status. Regarding the last resort use of LAWS, elevating them to the level of nuclear weapons should be done with caution, since it might exempt LAWS from all regulations. Accepting MHC for the development and use of LAWS would assure increasing autonomy in weapons systems does not create another mutually assured destruction deterrence structure already in place by nuclear weapons. MHC instead elevates LAWS above nuclear weapons due to the additional human involvement required even in last-resort situations.

Being exceptional, however, does not mean LAWS and MHC are mutually exclusive. The use of LAWS, if proven to be capable of adhering to the IHL, only moves the requirement of MHC up the command chain from the field operations. Here the direct impacts of human control are not significant, but the meaningfulness of the decision-making process leads to the activation and use of an autonomous weapons system. As LAWS is claimed to have revolutionary effects on warfare, new codes of conduct are needed regardless. Thus, MHC assures the kill chains are adequately re-evaluated for these new weapons systems.

By taking the direction proposed above, additional human control would not hamper the development of autonomous systems. It also pleases those worrying if international regulation has adverse effects on civilian AI development. Thus, a way for shared understanding and applicability of MHC in the treaty negotiations is made possible. Accepting autonomy to be developed and used but asking for more control for weapons systems in the kill chain makes MHC exceptional. It creates the need for additional clarification and reworked codes of conduct and operational principles for the military.

Following the argument put forward in this paper, the States Parties to the CCW focusing on practical steps such as articulating MHC in all parts of the command chain, creating explainable and understandable AI's, and avoiding creating loopholes such as last resort use in any regulatory texts on AI, are able to create new avenues for stronger regulation.

Bibliography

Aloyo, E. (2015), Just War Theory and the Last of Last Resort. *Ethics & International Affairs, 29*(2), 187-201.

Altmann, J. & Sauer, F. (2017), Autonomous Weapon Systems and Strategic Stability. *Survival, 59(5)*, 117-142.

Amoroso, D. & Tamburrini, G. (2021), In Search of the 'Human Element': International Debates on Regulating Autonomous Weapons Systems. *Italian Journal of International Relations, 56*(1), 20-38.

Article 36 (2014), *Key Areas for Debate on Autonomous Weapons Systems*. Available <

BDI (2021), *Comments on the proposal for A European AI Regulation*. Available < 20211108 Position BDI Proposal for a Regulation on Artificial Intelligence.pdf>

Bode, I. & Watts, T (2021), *Meaning-less human control*. University of Southern Denmark.

Campaign to Stop Killer Robots' 2020 publication on Frequently Asked Questions on Key Elements of a Treaty on Fully Autonomous Weapons

Cottrell, P. M. (2009), Legitimacy and institutional replacement: the convention on certain conventional Weapons and the emergence of the mine ban treaty. *International Organization, 63*(2), 217-248.

Crootof, R. (2016), War Torts: Accountability for Autonomous Weapons. *University of Pennsylvania Law Review, 164(6),* 1347-1402.

Cummings, M. L. (2004), Automation Bias in Intelligent Time Critical Decision Support Systems. *AIAA 1st Intelligent Systems Technical Conference 20-22 September 2004.*

Decision of Finland's Deputy Ombudsman (2018), EOAK/3379/2018. Accessible <https://www.oikeusasiamies.fi/r/fi/ratkaisut/-/eoar/3379/2018 >

DeepMind (2021), AlphaGo – The story so far. Accessed 13.10.2021. <https://deepmind.com/research/case-studies/alphago-the-story-so-far >

Department of Defence (2018), *Summary of the 2018 Defense Artificial Intelligence Strategy.*

Dilman, I. (1999), *Free will: An historical and philosophical Introduction*. Routledge.

European Commission (2019), *Ethics Guidelines for Trustworthy AI.*

European Commission (2020), *White Paper on Artificial Intelligence – A European Approach to excellence and trust*. COM(2020) 65 Final.

European Parliament (2021), *Resolution on Artificial intelligence*. 2020/2013 (INI).

Ford, C. (2017), Autonomous Weapons and International Law. *South Carolina Law Review, 69*(2), 413-.

Geist, E. M. (2016), It's already too late to stop the AI arms race – We must manage it instead. *Bulletin of the Atomic Scientist, 72*(5), 318-321.

Henriksen, A. & Ringsmose, J. (2015), Drone Warfare and morality in riskless war. *Global Affairs, 1,* 1-7.

Hoffman, F. (2018), Squaring Clausewitz's Trinity in the Age of Autonomous Weapons. *Orbis (Philadelphia), 63*(1).

Holland, M. A. (2020), *The Black Box, Unlocked: Predictability and Understandability in Military AI.* Geneva, Switzerland: United Nations Institute for Disarmament.

Horowitz, M. & Scharre, P. (2015), *Meaningful Human Control in Weapon Systems: A Primer*. Center for a New American Security. Working Paper.

Human Rights Watch (2015), *Ming the Gap: The Lack of Accountability for Killer Robots.*

Human Rights Watch (2020), *Stopping Killer Robots: Country Positions on Banning Fully Autonomous Weapons and Retaining Human Control.*

Human Rights Watch (2021), *Crunch Time on Killer Robots.*

ICJ (1996), Advisory opinion on the Legality of the Threat or Use of Nuclear Weapons.

ICRC (2021), IHL Database – Customary IHL. Accessible < https://ihl-databases.icrc.org/customary-ihl/eng/docindex/v1_rul_rule71>

ICRC (2021), *Position on Autonomous Weapons Systems.*

Israel's statement during the 1st meeting of the 3rd session in the CCW GGE 2021.

Kahn, P. (2002), Paradox of riskless war, *Philosophy and Public Policy Quarterly*, *22*(3), 2-8.

Kovic, M. (2018), *The Strategic paradox of autonomous weapons.* ZIPAR Policy Brief. Zurich, Switzerland.

Larson, E. J. (2021), *The Myth of Artificial Intelligence – Why Computers Can't Think the Way We Do.*

Lee, Kai-Fu (2018), *AI Superpowers: China, Silicon Valley, and the New World Order*. Boston: Houghton Mifflin Harcourt.

Marquelis, P. (2016), Making Autonomous Weapons Accountable: Command Responsibility for Computer-Guided Lethal Force in Armed Conflicts. *In Research Handbook on Remote Warfare*, ed. Jens David Ohlin.

McDougall, C. (2019), Autonomous Weapon Systems and accountability: Putting the cart before the horse. *Melbourne Journal of International Law, 20*(1), 1-30.

McIntosh, S. E. (2011), The Wingman-philosopher of MiG Alley: John Boyd and the OODA loop. *Air power history 58(4)*, 24-.

Methani, L., Tubella, A., Dignum, V. & Theodorou, A. (2021), Let me take over: variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence, 4:737072.*

Pattison, J. (2015), The ethics of diplomatic criticism: The Responsibility to Protect, Just War Theory and Presumptive Last Resort. *European Journal of International Relations, 21*(4), 935-957.

Reaching Critical Will (2021 a), Less Autonomy: More Humanity. CCW *Report, 9*(10). Available < https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2021/RevCon/reports/CCWR9.10.pdf>

Reaching Critical Will (2021 b), *Fact Sheet on Fully Autonomous Weapons*. Accessed 9.12.2021. < https://www.reachingcriticalwill.org/resources/fact-sheets/critical-issues/7972-fully-autonomous-weapons >

Roff, H. (2016), To ban or regulate autonomous weapons: A US response. *Bulletin of the Atomic Scientists, 72*(2), 122-124.

Roff, H. (2019), The Frame problem: The AI "arms race" isn't one. *Bulletin of the Atomic Scientists, 75*(3), 95-98.

Rosert, E. & Sauer, F. (2019), Prohibiting Autonomous Weapons: Put Human Dignity First. *Global Policy, 10*(3), 370-375.

Santoni De Sio, F. S. & Van den Hoven, J. (2018), Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI, 5:15.*

Scharre, P. (2019), *Killer apps: The Real Dangers of an AI arms race*. Foreign Affairs. Accessed 6.10.2020. Available < https://www.foreignaffairs.com/articles/2019-04-16/killer-apps>

Scharre, P. (2021), Debunking the AI Arms Race. *Texas National Security Review, 4*(3), 121-123.

Scharre, Paul (2018), "Army of None: Autonomous weapons and the future of war", *National Defense, 102(771).*

Simpson, T. W. & Müller, V. C. (2016), Just Wars and Robots' Killings. *The Philosophical Quarterly, 66*(263), 302-322.

Singer, Peter (2009), *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. New York: Penguin Press The term "meaningful human control" was introduced in 2014 by Article36 in their briefing paper *Key Areas for Debate on Autonomous Weapons Systems.*

Sparrow, R. (2016), Robots and Respect: Assessing the Case Against Autonomous Weapon Systsems. *Ethics & International Affairs, 30*(1), 93-116.

UK's statement during the 1st meeting of the 3rd meeting in the CCW GGE 2021.

UN (2017), Autonomy in Weapon Systems. *Working Paper submitted by the United States of America.* CCW/GGE.1/2017/WP.6.

UN (2018), Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapon Systems. *Working Paper Submitted by the United States.* CCW/GGE.2/2018/WP.4.

UN (2018), 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.

UN (2019), Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. CCW/MSP/2019/9.

UN (2021), Considerations for the report of the Group of Governmental Experts of High Contracting Parties to the Convention on Certain Conventional Weapons on emerging technologies in the area of Lethal Autonomous Weapons Systems on the outcomes of the work undertaken in 2017-2021. *Working paper submitted by Russia.* CCW/GGE.1/2021/WP.1.

UNESCO (2021), *Report of the Social and Human Sciences Commission.* 41 C/73.

UNIDIR (2018), Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies. A Primes. *UNIDIR Resources No. 9.*

UNIDIR (2020), *The Human Element in decisions about the use of force.*

US statement in the 4th meeting of the 3rd session of the CCW GGE 2021

US statements to the CCW 10.4.2018; 27.3.2019; 26.3.2019, 28.8.2018; 9.4.2018.

Wagner, B. (2021), Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet, 11*(1), 104-122.

Walzer, M. (1977), *Just and Unjust Wars.* 5th ed. Basic Books: New York.

Walzer, M. (2004), *Arguing About War.* Yale University Press.

White House (2019), *Executive Order on Maintaining American Lea*