Focus on the Human-Machine Relations in LAWS

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Empfohlene Zitierung / Suggested Citation:
International Panel on the Regulation of Autonomous Weapons (iPRAW). (2018). *Focus on the Human-Machine Relations in LAWS.* ("Focus on" Report, 3). . https://nbn-resolving.org/urn:nbn:de:0168-ssoar-77406-2

Focus on the Human-Machine Relation in LAWS

"Focus on" Report No. 3

March 2018

International Panel on the Regulation of Autonomous Weapons

International Panel on the Regulation of Autonomous Weapons (iPRAW)

coordinated by:

Stiftung Wissenschaft und Politik (SWP) – German Institute for International and Security Affairs
Ludwigkirchplatz 3-4
10719 Berlin, Germany

March 2018

www.ipraw.org
mail@ipraw.org

This project is financially supported by the German Federal Foreign Office.

About iPRAW

Setting and Objectives: The International Panel on the Regulation of Autonomous Weapons (iPRAW) was founded in March 2017. iPRAW is an independent group of experts from different nation states and scientific backgrounds. The panel will complete its work by the end of 2018.

The mission of iPRAW is to provide an independent source of information and consultation to the Group of Governmental Experts (GGE) within the framework of the United Nations CCW (Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects) during the ongoing process toward a possible future regulation of LAWS (Lethal Autonomous Weapon Systems). This work includes, but is not limited to, the provision of expertise on the military, technical, legal, and ethical basis for practical and achievable policy initiatives regarding LAWS. The mandate of the CCW's open-ended GGE on LAWS will guide the work of iPRAW.

iPRAW seeks to prepare, support, and foster a frank and productive exchange among participants, culminating in perspectives on working definitions and recommendations on a potential regulation of LAWS for the CCW GGE. iPRAW is independent from the GGE and does not function in any official capacity regarding the CCW.

Funding, Organization, and Participants: iPRAW is financially supported by the German Federal Foreign Office. The views and findings of iPRAW do not reflect the official positions of the German government or any other government. Stiftung Wissenschaft und Politik – The German Institute for International and Security Affairs (SWP) and the Johns Hopkins University Applied Physics Laboratory (JHU APL) are jointly organizing the panel. The participants have been selected on the basis of their expertise and the perspectives they bring from a wide range of professional and regional contexts. iPRAW represents the diversity of views on the topic of autonomy in weapon systems. Its members have backgrounds in natural science, engineering, law, ethics, political science, and military operational analysis.

Scope: The panel acknowledges that LAWS may pose a number of considerable legal, ethical and operational challenges and that they might change the security environment in a fundamental way. The full potential of these weapon systems is yet unknown and a mutually agreed definition on LAWS does not exist.

In order to support the CCW GGE process, iPRAW will work on how LAWS should be defined as well as on suggesting possible approaches to regulation. The panel's working sessions will cover the following topics

- state of technology and operations as well as existing definitions of LAWS
- computational systems within the scope of LAWS
- autonomy and human control
- ethics, norms and public perception
- risks and opportunities
- IHL and other fields of law.

iPRAW will publish working documents on each of these topics and will, in addition, publish the panel's final recommendations aimed at informing the CCW process.

Procedure: The participants commit themselves to actively participate in and contribute to all meetings and the scientific dialogue in-between meetings. The panel will meet seven times over the course of two years, starting in March 2017. Each meeting will take two and a half days and will be hosted by SWP in Berlin. Papers with agreed upon recommendations on relevant issues will be drafted and published via the project's website (www.ipraw.org) in-between meetings.

Communication and Publication: The participants discuss under the Chatham House Rule: participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed. As a matter of confidentiality, photographs, video or audio recordings as well as all kinds of activities on social media are not allowed during iPRAW meetings.

The results of the panel discussions will be published. iPRAW members will strive to reach consensus on their recommendations and to reflect that in the panel's publications. Media inquiries with regard to official iPRAW positions should be directed to the steering group. Apart from that, the panel members are free to talk about their personal views on participation and the topics of the panel.

Learn more about iPRAW and its research topics on **www.ipraw.org**. Please direct your questions and remarks about the project to **mail@ipraw.org**.

CONTENT

FIGURES & TABLES

EXECUTIVE SUMMARY

The International Panel on the Regulation of Autonomous Weapons (iPRAW) is an independent, interdisciplinary group of scientists working on the issue of lethal autonomous weapon systems (LAWS). It aims to support the current debate within the UN Convention on Certain Conventional Weapons (CCW) with scientifically grounded information and recommendations for the potential regulation of LAWS. Defining LAWS is a critical element of the CCW debate and as such a major component of iPRAW's mission.

iPRAW publishes interim reports that each focus on different aspects of or perspectives on LAWS, finished by a comprehensive final report in late 2018. This report focuses on machine autonomy and human control and their relevance with regard to LAWS. Building on the observations stated in this report, iPRAW makes the following conclusion for aspects of a potential regulation of LAWS:

The panel concludes that the leading characteristic of the human-machine interaction should be that of human control and machine dependence on humans in the execution of the targeting cycle. The control exercised by the operator must be sufficient to reflect the operator's intention for the purpose of establishing the legal accountability and ethical responsibility for all ensuing acts.

Legal as well as operational considerations show the necessity of human control over weapon systems with the requirement for predictability through a strong human involvement as one common denominator. In addition to that, a philosophical perspective on control helps to define this abstract concept, showing that reliability and predictability determine the level of control that humans must ascertain over objects.

To allow for predictability and to abide by legal requirements, the human operator must be aware of the state of the system as well as its environment. Therefore, the system's design must allow the operator to monitor both. This can be achieved through frequent points of inquiry throughout the targeting cycle. In addition to this situational understanding, the human operator needs options to interact with the system. From that iPRAW derives three recommendations to CCW States Parties:

Recommendation 1: Attempts to define LAWS by categorizing the level of autonomy within the system are not helpful. Instead, iPRAW recommends focusing definitions and potential regulations on the human responsibility and its role in weapon systems. Human control over machines' actions is crucial for legal as well as operational reasons.

Recommendation 2: The design of weapon systems with autonomous functionalities must enable the operator to **understand** the operational situation to allow for informed decisions over the use of force. The necessary monitoring of the environment and the system includes system diagnostics, internal and external sensors for system and environmental monitoring as well as methods for communicating that information. In addition, the ability for humans to actively **intervene** prior to the ultimate use of force should be a default feature.

The need for situational understanding and intervention is not limited to one single weapon system, but should also refer to systems of multiple robots executing a mission, which is how these capabilities will be developed and fielded.

Recommendation 3: States should negotiate a protocol to the CCW to regulate LAWS and incorporate the standards that will be enunciated in the said protocol in their domestic laws. Accordingly, States must develop or modify policy, law and practices to ensure proper human accountability of designers, operators, teammates and commanders for the outcomes of using weapon systems with autonomous functionalities.

Future iPRAW reports will continue to examine LAWS along the lines of these considerations.

1 INTRODUCTION

The International Panel on the Regulation of Autonomous Weapons (iPRAW) is an independent, interdisciplinary group of scientists working on the issue of lethal autonomous weapon systems (LAWS).[1] It aims to support the current debate within the UN Convention on Certain Conventional Weapons (CCW) with scientifically grounded information and recommendations for the potential regulation of LAWS. Defining LAWS is a critical element of the CCW debate and as such a major component of iPRAW's mission.

iPRAW publishes interim reports that each focus on different aspects or perspectives on LAWS.[2] This report focuses on the relation of machine autonomy and human control with regard to LAWS. Building on the observations stated in this report, iPRAW makes the following conclusion for aspects of a potential regulation of LAWS:

iPRAW chose 'human control' as the category to analyze the human-machine relation. By 'autonomy' we mostly refer to an attribute of weapon systems (i.e. autonomous functions), assuming that those are (and will be) the crucial problem to be confronted by global regulations.

The panel concludes that the leading characteristic of the human-machine interaction should be that of human control and machine dependence on humans in the execution of the targeting cycle. The control exercised by the operator must be sufficient to reflect the operator's intention for the purpose of establishing the legal accountability and ethical responsibility for all ensuing acts.

This report examines minimum requirements for human control over weapon systems derived from international legal principles as well as operational necessities. To that

1 While focusing on *lethal* autonomous weapon systems, iPRAW does not exclude a regulation of non-lethal force.

2 This particular report is based on the fourth meeting of iPRAW ("Autonomy and Human Control") in January 2018. The panel thanks **Julia Buchholtz** and **Frank Flemisch** for their valuable contributions to the meeting.

end, iPRAW regards human control as a constant process throughout the targeting cycle.

Autonomy is a notoriously difficult concept to define and delineate. Within the context of the debate on LAWS, whether a weapon enjoys autonomy or not is probably the most vexing issue. Tied into the question of machine autonomy is the intertwined concept of human control. These two terms exist on the same conceptual sliding scale – more machine autonomy means less human control, and more human control necessarily translates into less machine autonomy. This relationship underscores the importance of discussing these two terms jointly.

With this in mind, the panel believes that while it is helpful to have a clear *understanding* of autonomy for the purposes of any discussion on LAWS, it is not necessary to *define* the term for the purposes of any debate on regulations. Instead, engaging the issue from the control side of the sliding scale is a promising and productive endeavor.

Control is the key term to underpin any possible regulation and future debates on this issue. The question of control brings different aspects of the human-machine relation to the fore. At the same time it is essential to understand that control is a context-dependent term. This creates a need for probing varying possible implementations of control and how they impact the human-machine relation.

The international debate about LAWS references a number of different concepts of this relation already. They go under names such as human involvement[3], human judgement[4] or, most prominently, meaningful human control[5]. iPRAW acknowledges these concepts as valuable contributions to the debate and adopts a complementary perspective on the term control to enhance the current evolving understanding of human control and, in addition, to facilitate an understanding of operational consequences for the use of weapon systems with autonomous functions. The necessity for human control in the use of force (i.e. over autonomous functions in weapon systems) is an evolving international norm, which is underscored by reiterations of the concept by various CCW States Parties.[6] Therefore, there is ample reason to believe that human control is a valid concept to be fleshed out further and potentially become accepted practice later on.

[3] See CCW (2016), *Recommendations to the 2016 Review Conference – Advanced Version, Submitted by the Chairperson of the Informal Meeting of Experts.*

[4] See e.g. United States Department of Defense, *Directive 3000.09.*

[5] Heather Roff & Richard Moyes, (April 2016), *Meaningful Human Control, Artificial Intelligence and Autonomous Weapons*, Article 36.

[6] For more details about a potential codification of 'meaningful human control' as an emerging international norm see Elvira Rosert (2017), *How to Regulate Autonomous Weapons. Steps to Codify Meaningful Human Control as a Principle of International Humanitarian Law.*

2 The Human-Machine Relation

The following chapter touches upon several perspectives relevant to the assessment of the human-machine relation in weapon systems based on an interdisciplinary approach. First, we highlight autonomous functions in their relation to human control, especially in case of failure. Second, we assess the legal necessity for human control (possibly as an international norm) resulting from inadequacies of existing legal approaches. The third part deals with the operational perspective on the incentives and constraints of human control in the targeting cycle, followed by philosophical considerations about the nature of control.

2.1 Autonomy as a Function Requiring Control

In order to illumine the legal, military, and philosophical necessity to maintain control over autonomous functions in weapon systems, it is important to articulate the state of machine autonomy and its issues.

To date, despite the high-paced innovation in robotics and computational methods,[7] we do not yet have machines that operate in a truly autonomous capacity without any human input. Rather than a single linear spectrum of autonomy from human to machine, technologists at the frontiers of innovation discuss **autonomy as a spectrum with respect to specific functions**. A drone may be fully autonomous in navigating between two places (a function), yet it may fully rely on humans to determine and command which location is safe for it to land (another function). Most technologies of today have some form of autonomy in specific, defined sets of functions, but never in all of its functionalities. This implies that machines will, at least for the foreseeable future, require interaction with and

iPRAW's previous report on computational methods in the context of LAWS elaborates on the technology behind autonomous functions. One crucial finding was that the **unique judgement** of human decision makers cannot be replaced due to the inherent limitations of computational methods.

[7] iPRAW uses the term "computational methods" when it refers to techniques like "Artificial Intelligence" and "machine learning", as those imply a deeper meaning behind the logical and statistical methods and may lead to the false impression of the machine's intention and purpose.

reliance on human capacities and input, to varying degrees for varying functions that constitute a task.

Given the high stakes of failure in such a system, it is imperative that to design the system to **fail gracefully** and along defined procedures when it is no longer able to perform its autonomous functions. For some critical functions it even needs to be designed in such a manner that the specific function can be fulfilled only by handing over the control to a human being. To allow for a handing-over procedure, a robotic system needs to have defined and comprehensible modes of autonomous functionalities. Otherwise, humans will not be able to regain control in (time) critical situations as they do not understand the system's operational mode and will lack contextual awareness. Control is thus a necessity for the implementation of the human-machine-handshake protocol. Since even well designed failure modes bear a certain risk, the choice of autonomous functions needs to be considered carefully in the first place. Migrating from a mode A to a higher mode of autonomy B should be subject to direct human intervention: Cases where a person fails to perform critical functions of a mission and the machine takes over (similar to the anti-lock system in cars) have to be designed very carefully to avoid undermining this principle.

Consequently, rather than defining autonomy in an abstract sense, within the LAWS debate, the more practical approach is to adopt the concept of autonomy as a notion that is tied to a specific function of a system and different states of how autonomy drives these functions, rather than an inherent feature or quality of a weapon system.

2.2 The Legal Necessity for Human Control in Weapon Systems

States have a legal responsibility regarding the creation of new weapons under **Article 36** of the 1977 Additional Protocol to the Geneva Conventions, which states:

> "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".

Article 36 stipulates that states conduct reviews of new weapons to assess compliance with IHL. However, Article 36 cannot be considered a sufficient compliance method as only a handful of states carry out weapons reviews regularly and no international consensus exists on how such a review could fulfill a sufficient criteria for regulation. Even the states that carry out weapons reviews do so in a manner that is not considered transparent enough. Additionally, there are challenges with the testing and evaluation of autonomy that may render the review process even more difficult.

Compliance-based approaches often based on Article 36 observe the outcome of the use of force to judge whether the weapon system is (used) IHL-conformant. However, they do not prescribe the way to implement the steps necessary to comply with IHL. But as learning systems, by definition, learn and adapt to new environments, States can hardly ascertain that new weapon systems will reliably (and continuously) abide by the rules of international law.

With that in mind, human control is relevant as a principle for two reasons. First, it is relevant due to customary obligations to uphold the core rules of IHL, namely distinction, proportionality and the choice of weapon. At this point in time, it is hard to determine whether autonomous systems will be able to comply with these rules. Therefore, precaution, another rule of IHL, would be applicable: The **precautionary principle** requires that actors take actions to prevent harm.[8] Human control as a relevant precautionary principle would be essential. The second reason is based on Human Rights Law. The foundational right pertinent to this branch of international law is **dignity**. Eventually, if autonomous systems were to make life-and-death decisions, they could be non-compliant with this basic human right to dignity; an aspect iPRAW will further assess in a subsequent report.

2.3 Operational Perspectives on Human Control

In addition to legal requirements on control of weapons that a state chooses to use, militaries and fighting forces have multiple incentives and constraints for maintaining control over their capabilities. These incentives and constraints range from the philosophical to the pragmatic, which makes them applicable across a wide range of nations and actors. iPRAW believes that understanding the operational military perspective on the employment of force is important in any analysis of LAWS to understand the overlap of operational, political, legal, and ethical incentives as well as constraints.

At a very pragmatic level there is a strong incentive to prevent fratricide amongst military forces – both personnel and equipment – as a way to maintain combat power and to reduce risk to friendly forces. This overlaps with another pragmatic desire to achieve operational effectiveness, which is based on precision, predictability, and lethal efficiency. At a high, but still pragmatic level, political leaders seek to maintain control over escalation. At a higher philosophical level those states that are part of the international community typically seek to conduct themselves within the limits of their national laws and political structures, which includes international law and treaty obligations. Conduct within acceptable international norms and national cultures also guides a military's use of force within the context of control.

The dynamic targeting cycle as used by the U.S. armed forces consists of six steps: find, fix, track, target, engage, and assess.

Militaries utilize a range of mechanisms to achieve and maintain control over violence. Advanced weapon systems have multiple modes, safety mechanisms and highly scripted procedures for use that are developed to lead to successful engagements and to prevent fratricide at the same time. Extensive planning occurs in deliberate and dynamic targeting cycles to ensure the most effective weapons are used in a way that prevents fratricide, achieves operational objectives, and complies with the legal justification and political basis for conducting operations. The concept of command, which is vested with culture, authority, responsibility and accountability to maintain control of violence in the pursuit of national objectives, is abstract, but crucial to maintaining control. The hierarchies of the chain of command, the delegation or reserve of crucial decisions and the vast enterprise of staff, communications and information systems are all based on the need to maintain control. Many states'

[8] iPRAW that aspect in more detail in iPRAW (November 2017), *Focus on Computational Methods in the Context of LAWS (Report No. 2)*, pp. 17-18.

militaries have legal officers whose main role is to advise on the legality of particular aspects of operations and this includes a role in targeting cycles.

Targeting cycles, which have been discussed in previous iPRAW 'Focus on Reports',[9] are the procedures that many advanced military forces use to conduct deliberate (or planned) and dynamic targeting against their adversaries. **Deliberate targeting cycles** are really planning processes involving the selection of sets of targets to be attacked by pre-planned missions. These cycles often take long periods of time and involve targeting decisions for an entire operation. **Dynamic targeting cycles** are intended for time critical targets that are determined ahead of time to be priorities. The basic steps of the process are the same, but they occur more rapidly. There are variations between countries and alliances, but almost all of the recognized processes follow some form of finding targets, making determinations about them, engaging them with appropriate weapons and then assessing the effects. The steps of the cycle are intended to efficiently involve the appropriate decision-makers and maximize the military effectiveness of targeting while at the same time maintaining control over the use of force.

However, the development, fielding and eventual use of weapon systems and other military capabilities that utilize computational methods present a challenge to these traditional and time-tested methods of controlling the use of force by militaries. **Designing novel modes of control into systems is of crucial significance to maintain control and to provide a "failsafe" against a loss thereof.** Additionally, the design of interfaces and methods to monitor machine decision points during offensive operations is critical to enable humans to exercise the appropriate amount of oversight and control.

2.4 A Philosophical Concept of Control

Control and responsibility are inextricably linked. In general, people shall not be responsible for events that are "beyond their control", meaning the outcomes or risks that were not reasonably foreseeable. This notion has come to characterize many laws across the globe. There are circumstances where control is both an ethical and legal requirement of particular persons, and if control is absent under those circumstances the person or persons concerned may be deemed culpable. A clear conception of control is therefore vital. The philosopher and cognitive scientist Daniel Dennett defines control as follows:

> **"A controls B if and only if the relations between A and B is such that A can drive B into whichever of B's normal range of states A wants B to be in."**[10]

Put simply, the prerequisites for an agent to have control over something are that the agent a) has desires about the state or behavior of that thing, and b) the ability to cause the thing in question to go into the desired state or to behave in the desired way. A corollary of the second prerequisite is that the agent must have the knowledge necessary to be able to affect the state or behavior of the thing that is being controlled. Control is never *absolute* for non-omnipotent beings – there are

[9] See iPRAW (August 2017), *Focus on Technology and Application of Autonomous Weapons (Report No. 1)*, p. 12 and iPRAW *Report No. 2*, pp. 14-15.

[10] Daniel C. Dennett (1984), *Elbow Room. The Varieties of Free Will Worth Wanting*, p. 52.

always some factors that cannot be controlled. This does not, however, remove the possibility of *effective* control. As Dennett points out, "Foreknowledge is what permits control."[11] That is, an agent does not need to be in control of every factor in a particular environment in order to be able to sufficiently control something in that environment, as long as the agent can reliably predict the effect each relevant factor will have on the object in question. For example, a human throwing a stone cannot control the earth's gravitational pull, but her ability to predict the effects that gravity will have on the stone once it is thrown gives her, all other things being equal, sufficient control over her projectile. *Predictability* is therefore critical to control. The more predictable the environment and the more predictable the behavior of the object of control, the lower the epistemic threshold for effective control. The objective however, shall be *reliable*; it must repeatedly fulfill the expectations set upon its performance. **Reliability and predictability determine the level of control that humans must ascertain over objects.**

An important implication of this concept of control is that control and direct manipulation are not necessarily synonymous. For example, a modern car using drive-by-wire technology such as electronic stability control, in which the traditional mechanical controls of the vehicle are replaced by a computerized electronic system, reduces the directness of the driver's manipulation of the car's wheels, brakes, and the like. Through this, the assisting system gives him more control over the car's path and the outcome of the situation.[12] Similarly, a precision guided firearm[13] increases the shooter's ability to cause the bullet to hit the desired target (i.e., following the definition above, it increases the shooter's degree of control), but does so via a mechanism which decreases the shooter's direct manipulation of the weapon's aiming system. It follows from this that the directness of the means whereby the agent seeks to control some object is only contingently related to the degree of control. Under some circumstances more direct manipulation 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.

The implications for the debate over LAWS that emerge from this philosophical perspective on the concept of control include the following: It is important that discussions on the notion of control in the context of LAWS do not erroneously confuse 'control' with 'direct manipulation'. Assisting systems can be helpful and still allow for control over the machine's actions. **However, as technology stands, and given the capacity of the human-machine team to adapt appropriately to changes in its operating environment, the need for predictability dictates that effective control requires a situational understanding of the human operator.**

11 See Dennett 1984, p. 54.

12 At the same time, the assisting system reduces the options of the driver. If he actually wanted to crash the car/leave the road for whatever reason, he would not be able to do so. On that regard, the assisting systems and the lack of direct manipulation would reduce his control over the car and thus the outcome of the situation.

13 A firearm which incorporates a ballistic computer that adjusts the weapon's point of aim to compensate for factors like the bullet's ballistic trajectory and the effect of wind velocity and direction.

3 Requirements for Control in Tasking and Execution

Humans could have different 'relationships' with LAWS, as a human could be a commander, operator, teammate, legitimate target or civilian. Each of these roles entails different types of interaction, ranging from control to involvement to oversight or mere reactions to the weapon's attack or the system's instructions. iPRAW is primarily concerned with the analysis of human control over weapon systems, which is why this section focuses on humans in roles such as operator, teammate or military commander. This chapter explores requirements for control by those human roles.

3.1 Modes of Control

Humans exercise control in different ways and at different stages – through design choices (e.g. via explainable AI, testing, certification processes), mission tasking, and mission execution (e.g. compliance with IHL). Deliberate design choices are crucial for enabling operators, teammates and commanders to maintain the desired and appropriate level of control during tasking and execution.

By 'control by design' iPRAW refers to the technical specifications of the system like the interface. It is a necessary condition for 'control in use', which addresses the operational dimension of control.

Control by design focusses on the technical requirements. It refers to a specific hard- and software design which allows an operator to actually exercise control during the operation of the system. 'Control by design' calls for specific instruments in the human-machine interface and relevant procedures programmed into the system's processes to enable human input and intervention. The concept of 'control by design' encompasses the possibility to set 'probes' to illustrate and assess the system's state and actions in the various steps of the targeting cycle. It is a necessary condition for **control in use**, which encompasses the procedural requirements to maintain control over the systems during planning, tasking and operation.

Combinations of both modes are physical design choices that, amongst others, create deliberate technical limitations on range or effect as ways to maintain control. For instance, many nations adhere to the range limitations on theater ballistic missiles

stipulated by the Intermediate Nuclear Forces (INF) treaty. Thus, this legal standard limits the effective range of these missiles, which could be designed to go further, in a way that exercises control by artificially limiting the ability of tactical and operational commanders. Additionally designing different modes of operation that must be deliberately set by humans during operations is another way to maintain control both over technology and subordinate personnel. For example, the AEGIS missile system used by the US and other navies has a very high degree of autonomous capability to track, prioritize and engage aerial targets – but it also has different modes of operation that feature very different limits on the ability of the technology or conversely the role of the sailors in engaging targets.

3.2 Situational Understanding and Intervention

Maintaining appropriate and desired control during tasking of a future LAWS and the LAWS executing its mission would require a two-step approach to system design. It consists of (1) the ability of the human to understand the situation and its context including the state of the weapon systems as well as the environment, and (2) the option to appropriately intervene if necessary.

Situational Understanding: Situational understanding means that the human operator is aware of the environment and the mode of the system during the operation. The continuous awareness regarding the environment is necessary because battlespace situations change, for instance if civilians enter or if a combatant surrenders or is wounded and thus *hors de combat*. The supervision of the system itself is important to discover malfunctions or hacking before a catastrophic effect occurs.[14] This also influences both the system's design and interface, as it must present the operator with that information.

Situational understanding as a dynamic concept could depend on context and application and may vary in quality and quantity even within a given system. It is therefore important to have minimum requirements for how humans can maintain situational understanding. On the one hand, these criteria should reflect the general unpredictability of military operations and the respective environments. On the other hand, they should enable the operator **to assess the computational outcome** of a task delegated to a machine. Such criteria are important to keep weapon systems from becoming black boxes and e.g. choosing wrong paths to find the correct solution. One way to approach this could be the common criteria to test mathematical solutions: Those must be plausible, relevant, consistent, complete or unique (for details see Annex 5.1).

The practical applicability of these criteria depends on the level of abstraction and the design choices. The operator does not have to be able to answer the questions on a software level, for example. Nevertheless, the system's design and interface must allow the operator to understand why the system has produced a specific outcome. That includes a basic understanding for the underlying data and data fusion as well

[14] See UNIDIR 2017, *The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations*, pp. 12-13.

as the logical principle (e.g. probabilities). Projects like *Explainable AI*, if successful, could offer a way to approach this issue.[15]

The design of failure modes in all stages of the targeting cycle must allow for enough time and information for situational understanding. That would include a clear indication of responsibilities (What is demanded from the operator? What can the machine still do on its own?) and an immediate halt on the use of force.

Intervention: To allow for actual control over the system's actions, the human operator must be able to override the system, meaning that the human should be able to appropriately manipulate the machine at any point in time. In a maximalist version, that would be realized in all steps of the targeting cycle – but at least during the target selection and engagement the human should intervene.

Examples of automated defensive weapon systems exist in which the interface of the system asks the operator for an active decision and action to initiate the use of force, whereas others offer a veto power. In that case, the system translates omission as an implied consent for the use of force. The first option slows down the operation considerably, the latter one imposes on the human to adapt to machine speed limiting the chance for situational understanding.

As machine speed is incomprehensible to humans already, iPRAW cautions against adapting the latter procedures in future weapon systems and expanding their use outside the one exception which is the very limited context of solely defending at incoming munitions or materiel.

3.3 Minimum Requirements for Control

Looking at a spectrum of options for human involvement, the different grades vary in their military (dis-)advantages and room for error. The following options are just a few examples, other combinations of (different levels of) situational understanding and intervention are possible.

frequent feedback for situational understanding	frequent feedback for situational understanding	reduced situational understanding through occasional updates	no immediate situational understanding
&	&	&	&
frequent option for intervention with mandatory human action to initiate the use of force	frequent option for intervention without necessary human action to initiate the use of force (veto option)	(machine initiated) deliberate human action prior to the use of force	no immediate option for intervention
highest level of human involvement (control)	precautionary resolution; supervision at human speed	fighting at machine speed with *ultimate human decision*	lowest level of human involvement, e.g. 'boxed autonomy'; fighting at machine speed

Figure 1: Options for Human Involvement

[15] See David Gunning, *Explainable Artificial Intelligence (XAI)*, DARPA; also iPRAW, *Report No. 2*, p. 12.

The **maximum** of control would require at least frequent communication[16] between the human operator and the machine to ensure the operator's situational understanding in combination with a deliberate intervention regarding the use of force. In that case, control over the use of force requires an active human operator in the targeting process as opposed to precautionary resolutions. That does not necessarily exclude swarm functionality, but would put a high bar to it. In general, the decision to the use of force may not be taken solely in advance but must occur within the operational situation. As fleshed out in iPRAW's previous report, frequent information between the steps in the targeting cycle would be necessary to avoid the accumulation of errors and uncertainties.[17]

Degrading steps could be **(1)** frequent situational understanding, but no necessary action by a human, possibly leaving a veto option or **(2)** no continuous situational understanding, but (machine initiated) deliberate human action regarding the use of force. The second option would leave the *ultimate* decision on the use of force to a human, but does not allow for a substantial situational understanding: the operator has to respond to the situation as filtered and presented by the system at a very high speed.

The **minimal version of human involvement** would include a 'boxed autonomy'[18] that is based on precautionary resolutions without an immediate situational understanding. All safeguards and information gathering would have been done in advance; the machine would follow predefined failure modes in case it detects problems by itself. In this version, a constant communication link would not be required. That allows for some advanced military applications in secluded areas, but it comes at a high price: there would be no option to detect and react in case of unpredictably changing environments or problems within the system. Those changes could be, for example, civilians, wounded or surrendering combatants or hacks of the system. Hacks and malfunctions can only be detected "from the inside", a mere optical surveillance via satellite pictures or videos by drones, for example, might not be enough. As with its previous reports, iPRAW remains critical about the concept of boxed autonomy.

The concept of 'boxed autonomy' consists of a predefined context in which the system has to locate and engage a target. The box conditions would be preprogrammed and combined with parameters limiting the system's abilities once it is outside of the range of communication.

16 For a variety of scenarios examining the role of communication on weapon systems with autonomous functions see iPRAW, *Report No. 1*, pp. 13-18.

17 See iPRAW *Report No. 2*, pp. 19-20.

18 See iPRAW *Report No. 1*, pp. 15-16 and *Report No. 2*, p. 18.

4 CONCLUSION

Understanding LAWS is a critical element of the CCW debate and is as such a major component of iPRAW's mission. As shown above, human control within the targeting cycle is necessary for legal, ethical and operational reasons. Although control is never absolute, it has to match the risk for infringements of IHL.

The 2017 GGE report on LAWS called for further consideration of the "human-machine interaction in the development, deployment and use"[19] of LAWS. **iPRAW considers that the leading characteristics of human-machine interactions should be human control and the system's dependence on humans in the execution of the targeting cycle. The control exercised by the operator must be sufficient to reflect their intention for the purpose of establishing the legal accountability and ethical responsibility for all ensuing acts.**

4.1 RECOMMENDATIONS

The concept of control over autonomous functions in weapon systems can be approached from different angles while situated in legal standards that are backed by international law. Above, we chose 'control by design' and 'control in use' as one analytical category; the distinction between situational understanding and intervention serves as a second one.

Both perspectives show that control cannot be defined universally, while merely context-dependent requirements would lead to arbitrary results. Therefore, minimum requirements for control should consider both elements.

[19] CCW (2017), *Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS) – Advanced Version*, p. 4.

	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

Table 1: Overview of Requirements for Control

Following the observations stated above, iPRAW makes the following recommendations for aspects of a potential regulation of LAWS:

Recommendation 1: Attempts to define LAWS by categorizing the level of autonomy within the system are not helpful. Instead, iPRAW recommends focusing definitions and potential regulations on the human responsibility and role in weapon systems. For legal as well as operational reasons, human control over the system's actions is crucial.

Recommendation 2: The design of weapon systems with autonomous functionalities must enable the operator to **understand** the operational situation to allow for informed decisions over the use of force. The necessary monitoring of the environment and the system includes system diagnostics, internal and external sensors for system and environmental monitoring as well as methods for communicating that information. In addition, the ability for humans to actively **intervene** prior to the ultimate use of force should be a default feature.

The need for situational understanding and intervention is not limited to one single weapon system, but should also refer to systems of multiple robots executing a mission, which is how these capabilities will be developed and fielded.

Recommendation 3: States should negotiate a protocol to the CCW to regulate LAWS and incorporate the standards that will be enunciated in the said protocol in their domestic laws. Accordingly, States must develop or modify policy, law and practices to ensure proper human accountability of designers, operators, teammates and commanders for the outcomes of using weapon systems with autonomous functionalities.

4.2 The Way Ahead

Future iPRAW sessions will continue to examine the role and impact of LAWS in warfare using additional methodologies. Upcoming reports will assess the role of **ethics** including human dignity, which is closely linked to international human rights law. This and other **legal frameworks** will be subject to further analysis, too, as extant legal rules may not suffice to address the present challenges: The current existing international norms and rules that are deemed to be universal may not satisfy the new scope of the problems posed by the use of AI-enabled weapons. Usually, laws focus on weapons and their use, but with regard to LAWS, the human-machine relation (and therefore the human role) adds another relevant element to the regulation. A new type or instrument of regulation might be necessary and iPRAW strives to address this issue more thoroughly in a subsequent report.

The question of **verification** is linked to the requirements for human control and autonomous functions as software based functions and interfaces can be difficult to retrace.

5 ANNEX

5.1 ADDITIONAL INFORMATION ON MATHEMATICAL TESTING

Full control over the execution of a task given to a machine represents the ability to test against these criteria at any given point before the result is being implemented. The concept of "solutions" is not technology-dependent (i.e. it does not require a specific interface or data format nor does it relate to a specific computational method or process) but rather requires to challenge the machine as a logical entity and its results as coherent logical processing. Or, in a nutshell, challenge the machine's solution as you would do it with a human decision. Following this concept (as it is done in math), a solution must meet the following criteria:

Plausibility: Is the solution plausible with regard to the given task and the environment? Is there a transparent logic leading from the input to the solution?

Relevance: Is the solution relevant for the given task? Mathematical models, which lay ground for most of the computational methods used in autonomous systems, often have solutions that are plausible and correct but are not relevant (i.e. trivial solutions). In complex models and calculations it may be difficult to distinguish those from relevant solutions.

Consistency/Reproducibility: Is the solution consistent with other solutions of the same or varying problems? Does sensor noise, for example, influence the outcome in an inconsistent way? In operational terms, this describes a reproducibility of the solution under a variation of stochastic change of the environment (noise) and should therefore safeguard predictability.

Completeness or Uniqueness: Is the solution or the set of solutions complete? Often tasks given to a computational system lead to a set of solutions. Without prioritization, all possible solutions must be represented in this set. If the solution is unique, is there a verification for it? Preferably, solutions to a given problem should be unique or highly probable to achieve (otherwise a prioritization is needed). In the real world, this is not always the case. In addition, a proof for uniqueness is a challenging task in theory, all the more in practice. It often requires alternative calculations models to check against (Do they have other solutions under the same condition?) and even these numerical methods often cannot prove a fundamental uniqueness of a specific solution to a given problem but rather lower the probability of finding alternative solutions.

5.2 LITERATURE

Referenced in the Report

CCW (2016), *Recommendations to the 2016 Review Conference – Advanced Version*, Submitted by the Chairperson of the Informal Meeting of Experts, <https://www.unog.ch/80256EDD006B8954/(httpAssets)/6BB8A498B0A12A03C1257FDB00382863/$file/Recommendations_LAWS_2016_AdvancedVersion+(4+paras)+.pdf> (March 26. 2018).

CCW (2017), *Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS) – Advanced Version*, <https://www.unog.ch/80256EDD006B8954/(httpAssets)/B5B99A4D2F8BADF4C12581DF0048E7D0/$file/2017_CCW_GGE.1_2017_CRP.1_Advanced_+corrected.pdf> (March 26, 2018).

Dennett, Daniel C. (1984), *Elbow Room. The Varieties of Free Will Worth Wanting.*

International Panel on the Regulation of Autonomous Weapons – iPRAW (August 2017), *Focus on Technology and Application of Autonomous Weapons (Report No. 1)*, <https://www.ipraw.org/wp-content/uploads/2017/08/2017-08-17_iPRAW_Focus-On-Report-1.pdf> (March 26, 2018).

Gunning, David, *Explainable Artificial Intelligence (XAI)*, DARPA, <https://www.darpa.mil/program/explainable-artificial-intelligence> (March 26, 2018).

International Panel on the Regulation of Autonomous Weapons – iPRAW (November 2017), *Focus on Computational Methods in the Context of LAWS (Report No. 2)*, <https://www.ipraw.org/wp-content/uploads/2017/11/2017-11-10_iPRAW_Focus-On-Report-2.pdf> (March 26, 2018).

Roff, Heather M.; Moyes, Richard, (April 2016), *Meaningful Human Control, Artificial Intelligence and Autonomous Weapons*, Article 36, <http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf> (March 26, 2018).

Rosert, Elvira (2017), *How to Regulate Autonomous Weapons. Steps to Codify Meaningful Human Control as a Principle of International Humanitarian Law*, <https://www.hsfk.de/fileadmin/HSFK/hsfk_publikationen/Spotlight0617.pdf> (March 26, 2018).

UNIDIR 2017, *The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations*, <http://www.unidir.org/files/publications/pdfs/autonomous-weapon-systems-and-cyber-operations-en-690.pdf> (March 26, 2018).

United States Department of Defense, *Directive 3000.09*.

Further Reading

Flemisch, F.; Schieben A., Schoemig, N., Strauss, M.; Lueke, S.; Heyden, A. (2011), *Designing Human Computer Interfaces for Highly Automated Vehicles: Issues under Consideration in the EU-Project HAVEit; International conference on Human Computer Interaction*, <http://www.springerlink.com/content/75044027h038h9g6/fulltext.pdf> (March 26, 2018).

Flemisch, F.; Heesen, M.; Hesse, T.; Kelsch, J.; Schieben, A.; Beller, J. (2011), *Towards a dynamic balance between humans and automation: Authority, Ability, Responsibility and Control in Shared and Cooperative Control Situations*; Int. Journal Cognition, Technology & Work <http://www.springerlink.com/content/x63032819hx560m3/> (March 26, 2018).

5.3 Members of iPRAW

Liran Antebi*
Research Fellow
Institute for National Security Studies
Tel Aviv, Israel

Peter Asaro
Professor
The New School
New York, USA

Deane-Peter Baker
Senior Lecturer
University of New South Wales
Canberra, Australia

Vincent Boulanin*
Researcher
Stockholm International Peace Research Institute
Stockholm, Sweden

Thompson Chengeta
Fellow
South African Research Chair in International Law, University of Johannesburg
Johannesburg, South Africa

Anja Dahlmann
Researcher
German Institute for International and Security Affairs
Berlin, Germany

Marcel Dickow
Head of Research Division
German Institute for International and Security Affairs
Berlin, Germany

Denise Garcia
Professor
Northeastern University
Boston, USA

Robin Geiß
Professor
University of Glasgow
Berlin, Germany

Erin Hahn*
Researcher
Johns Hopkins University Applied Physics Laboratory
Washington D.C., USA

Vadim Kozyulin
Researcher
PIR Center for Policy Studies
Moscow, Russia

Ian MacLeod
Researcher
Johns Hopkins University Applied Physics Laboratory
Washington D.C., USA

AJung Moon
Director
Open Roboethics Institute
Canada

Shashank Reddy
Researcher
Carnegie India
New Delhi, India

Heigo Sato
Professor
Takushoku University
Tokyo, Japan

Frank Sauer
Researcher
Universität der Bundeswehr
Munich, Germany

David Hyunchul Shim*
Professor
Korea Advanced Institute of Science and Technology
Daejeon, South Korea

Lena Strauß
Research Assistant
German Institute for International and Security Affairs
Berlin, Germany

The asterisk indicates those members, who did not participate in the fourth meeting of iPRAW ("Autonomy and Human Control") in January 2018.

Former members: **Dong Lin**, Researcher at the National University of Defense Technology in Changsha, China and **Kelvin Wong**, Researcher at IHS Janes in Singapore.

International Panel on the Regulation of Autonomous Weapons (iPRAW)

coordinated by:

Stiftung Wissenschaft und Politik (SWP) – German Institute for International and Security Affairs
Ludwigkirchplatz 3-4
10719 Berlin, Germany

March 2018

www.ipraw.org
mail@ipraw.org

This project is financially supported by the German Federal Foreign Office.

International Panel on the Regulation of Autonomous Weapons