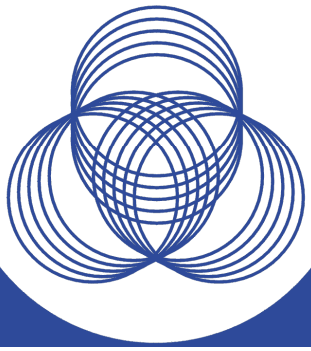
Food for Thought Paper



IFSH
IFAR

Autonomy of Weapon Systems

Christian Alwardt
Martin Krüger

FEBRUARY 2016

Institut für Friedensforschung und Sicherheitspolitik an der Universität Hamburg/
Institute for Peace Research and Security Policy at the University of Hamburg
Beim Schlump 83, 20144 Hamburg
Tel: +49 40 866 077-0 - Fax: +49 40 866 36 15

AUTONOMY OF WEAPON SYSTEMS

1. Context

The term *autonomy* is not new, but still requires a comprehensive analysis when applied to *weapon systems*. Immanuel Kant considered *autonomy* from a philosophic point of view as an individual's capacity for self-determination and consequently, as a crucial element to hold somebody morally accountable for their action.[1] This common, rather straightforward understanding of *autonomy* in the context of human action creates some challenges when applied to *machines* respectively *robots*[2]. *Robots* cannot be considered as *individuals*. They are (yet) not able to naturally develop a capacity of self-determination over time. It is a matter of fact that self-determining *robots* likewise human capacities do not exist so far, even though the research in this field has been dramatically progressing in recent years. Moreover, it is very hard to believe that in the near future *robots* will be considered accountable for their action as it goes for human beings. Despite these very obvious deficiencies, the term *autonomy* has been used lately more frequently in the context of *unmanned systems* in particular as military vocabulary. But it also has to be stated, not surprisingly, that there are numerous and quite different definitions and interpretations of this term in the literature. Recent discussions about *lethal autonomous weapons systems* (LAWS) at the Conference of the United Nations Convention on Conventional Weapons (CCW)[3] have shown that there is no common understanding of the term *autonomy*, including at international level. At the same time, a common understanding of *autonomy* is crucial in order to negotiate agreements on the prohibition or regulation of LAWS. In other words, how can something be controlled without a common understanding of the subject as such.

2. Objective

This paper shall open up new perspectives on the term *autonomy* when used to characterize *weapon systems* in order to better understand this term and feed the ongoing discussions related to this subject.

To this end, the paper will first analyze a variety of already existing definitions of *autonomy* by pointing out similarities and differences. Based on these findings, two different approaches will be applied then to classify *weapon systems* regarding their level of *automation*. For this, one approach addresses this from a two-dimensional and the other from a three-dimensional perspective. Finally, a summary of the outcomes will be presented.

3. An Analysis of Existing Definitions of Autonomy

Numerous definitions and categorizations of the term *autonomy* can be found in the literature. They are partially overlapping but they also address *autonomy* from different angles. Some regard *autonomy* as a fixed term and as a final result reached in an ongoing *automation process*. Others see it more as a spectrum described by certain, well-defined characteristics.

---

1 Cf. Kant, Immanuel: Grounding for the Metaphysics of Morals, Translated by Ellington, James W., Indianapolis/Cambridge, 1981.

2 The term *robot* is not clearly defined in the literature and is often used as a synonym for the term *machine*.

3 Convention on prohibitions or restrictions on the use of certain conventional weapons which may be deemed to be excessively injurious or to have indiscriminate effects, CCW.

The following analysis aims at showing these overlaps and differences and deriving new perspectives from these.

3.1 Autonomy and human control

One of the more complete known definitions of *autonomy* was published by Human Rights Watch in 2012. According to this model, the level of *autonomy* is defined by the degree of human control over specific tasks that the weapon system/ robot has to fulfill.

Model by Human Rights Watch*:*

- *Human-in-the-Loop Weapons: Robots that can select targets and deliver force only with a human command;*
- *Human-on-the-Loop Weapons: Robots that can select targets and deliver force under the oversight of a human operator who can override the robots' actions; and*
- *Human-out-of-the-Loop Weapons: Robots that are capable of selecting targets and delivering force without any human input or interaction.*[4]

(Full) autonomy is achieved when there is no human control over the whole process of selecting and engaging a target *(human-out-of-the-loop)*. With regard to this model the system is also categorized as *autonomous* even if the human operator has very limited oversight and can still intervene *(human-on-the-loop)*.

This understanding of autonomy is very similar to the definition provided by Paul Scharre and Michael C. Horowitz:

> „*autonomy is the ability of a machine to perform a task without human input [supervision]*."[5]

The model by Noel Sharkey is selected as another example. In this model the fifth level describes a fully autonomous system.

Model by Sharkey:

1. *Human engages with and selects target and initiates any attack*
2. *Program suggests alternative targets and human chooses which to attack*
3. *Program selects target and human must approve before attack*
4. *Program selects target and human has restricted time to veto*
5. *Program selects target and initiates attack without human involvement.*[6]

Numerous other models categorizing autonomy can be found in the literature besides the shown definitions above. All these definitions and categorizations have in common that they define the term *autonomy* by the degree of human control over specific information and decision processes.

However, this alone seems to be insufficient in order to define *autonomy* due to the missing distinction between systems that entirely operate on their own and can therefore be considered

---

4 Human Rights Watch/ IHRC: LOSING HUMANITY. The Case against Killer Robots, https://www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf (05.03.15), p. 2.

5 Schaare, Paul/ Horowitz, Michael C.: An Introduction to AUTONOMY in WEAPON SYSTEMS. Working Paper, February 2015, http://www.cnas.org/sites/default/files/publications-pdf/Ethical%20Autonomy%20Working%20Paper_021015_v02.pdf (02.03.15), p. 5.

6 Sharkey, Noel: The human control of weapons: a humanitarian perspective, https://www.law.upenn.edu/live/files/3948-sharkey---human-control-of-weapons-pf-draftpdf (05.03.15), p. 4.

as *autonomous* on the one hand and systems that simply follow (human) pre-programmed instructions and hence, have to be characterized as *automatic* on the other hand.

3.2 Autonomy and environmental complexity

A definition by Christof Heyns addresses the aspect of *autonomous* versus *automatic systems* by taking into account the complexity of the environment:

> "Automatic systems, such as household appliances, operate within a structured and predictable environment. Autonomous systems can function in an open environment, under unstructured and dynamic circumstances."[7]

With regard to this definition, *autonomous systems* even fulfill tasks independently on their own in changing, unpredictable environments, which *automatic* (pre-programmed) *systems* cannot do.

George Bekey presents a very similar interpretation of *autonomous systems*:

> "The capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time."[8]

Based on the previous analysis of various definitions and categorizations, two independent dimensions can be recognized interacting with and influencing a *system's autonomy*:

(1) Degree of human control over information and decision processes
(2) Complexity of operational environment

The processes performed by *systems* becoming more and more *automatic* or even *autonomous* can be understood as an ongoing *automation* of functions and tasks previously carried out by a human operator. Based on Raja Parasuraman et al., *automation* can be applied to four broad classes of functions: information acquisition, information analysis, decision and action selection, and action implementation.[9] Depending on the *weapon system's* technological capabilities, certain functions and tasks may be solely performed by the system without any human control, which leads to *partial autonomy*.

Keeping in mind the given *human control* and the *operational environment*, some questions still remain: Where do *automation* processes end and (full) *autonomy* is reached? How can different degrees of automation be distinguished? One thing is certain: One will not be able to draw a sharp line anyway.

In the following this paper presents two approaches dealing with the problem on how to define *autonomy* and distinguish different levels of *automation*. The first approach addresses this two-dimensionally by further scaling the two, already recognized, independent

---

7 Heyns, Christof: Report of the Special Rapporteur on Extrajudicial. Summary or Arbitrary Executions, A/HRC/23/47, New York: United Nations, 2013, p. 8.

8 Definition by Bekey, George in: Lin, Patrick/ Bekey, George/ Abney, Keith: Autonomous Military Robotics: Risk, Ethics, and Design, 2008, p. 103.

9 Parasuraman, Raja / Sherian, Thomas B. / Wickens, Christpher D. (2000): A Model for Types and Levels of Human Interaction with Automation, IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, Vol. 30, No. 3, May 2000.

dimensions *degree of human control* and *complexity of operational environment* and relating them to each other in order to define certain levels of *automatic* and *autonomy*. The second approach introduces an additional independent dimension: *system intelligence,* in order to allocate and classify *automated* and *autonomous weapon systems* within a three-dimensional space.

4. Autonomy – a two-dimensional approach

In order to define *autonomy and distinguish different levels of automation,* the two independent dimensions *degree of human control* and *complexity of operational environment* have to be characterized first. The chosen scale is based on the aforementioned definitions, but at the same time kept simplistic in order to give the model a universal character and thus broaden its applicability beyond *weapon systems*. The dimension *operational environment* is scaled by *predictable* and *unpredictable*. The dimension *human control over the information and decision processes* is scaled by *total*, *partial*, *supervised* and *none*. Although, many of the previously shown *autonomy* definitions and models do not distinguish between *no human control* and *supervision* due to the assumption that human oversight can only be poorly executed in practical terms, this paper keeps this distinction. This is justified by the fact that even recent international discussions at the CCW take into account *human supervision* by creating new terms like *meaningful human control*.[10] It is also an attempt to address in particular ethical and legal issues with regard to *autonomous systems*.

This model is built upon the assumption, that the consideration of those critical questions constitutes an inherent part of the information and decision processes and therefore does not exhibit an extra dimension here. Figure 1 schematically depicts the two dimensions of *autonomy* and relates them to each other.

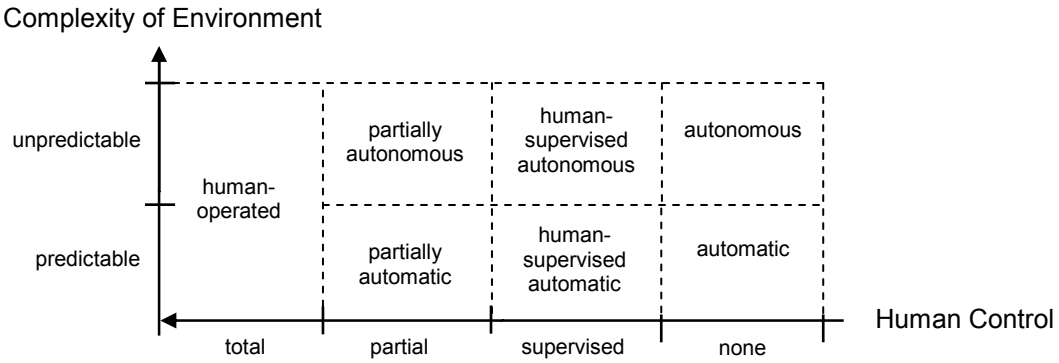


Figure 1: Two-dimensional approach to distinguish different levels of automation.

[10] The term "meaningful human control" played a povital role during the CCW-Conference 2014; cf. Knuckey, Sarah: Governments Conclude First (Ever) Debate on Autonomous Weapons: What Happened and What's Next, Justsecurity.org, 16.05.2014, http://justsecurity.org/10518/autonomous-weapons-intergovernmental-meeting/ (25.03.15).

The following definitions apply accordingly:

Level of Automation	Definition
human-operated	perform all of the information and decision processes only with a human input
partially automatic	perform parts of the information and decision processes without any human input in a predictable environment
human-supervised automatic	perform all or parts of the information and decision processes under human oversight in a predictable environment
automatic	perform all of the information and decision processes without any human input in a predictable environment
partially autonomous	perform parts of the information and decision processes without any human input in an unpredictable environment
human-supervised autonomous	perform all or parts of the information and decision processes under human oversight in an unpredictable environment
autonomous	perform all of the information and decision processes without any human input in an unpredictable environment

Table 1: Levels of automation.

5. Autonomy – a three-dimensional approach

Automated weapon systems are usually distinguished with regard to their level of automation, which has tended to be defined in different manners by means of discipline-specific characteristics. This three-dimensional approach is to distinguish automated weapon systems more generally from each other with regard to their system capabilities, their freedom to act on their own, and the type of environment in which they are intended to operate. A high degree of system intelligence does not necessarily result in overall system autonomy (which depends on the remaining human control). However, depending on the area of use (type of environment, special tasks etc.), overall system autonomy may demand a certain degree of system intelligence.

Therefore, a three-dimensional space is spanned by three *independent* vectors, namely

- ✦ "***System intelligence***" (abilities regarding sensing and information processing → situational-awareness)
- ✦ Intended level of **human control**
- ✦ Complexity of **operational environment**

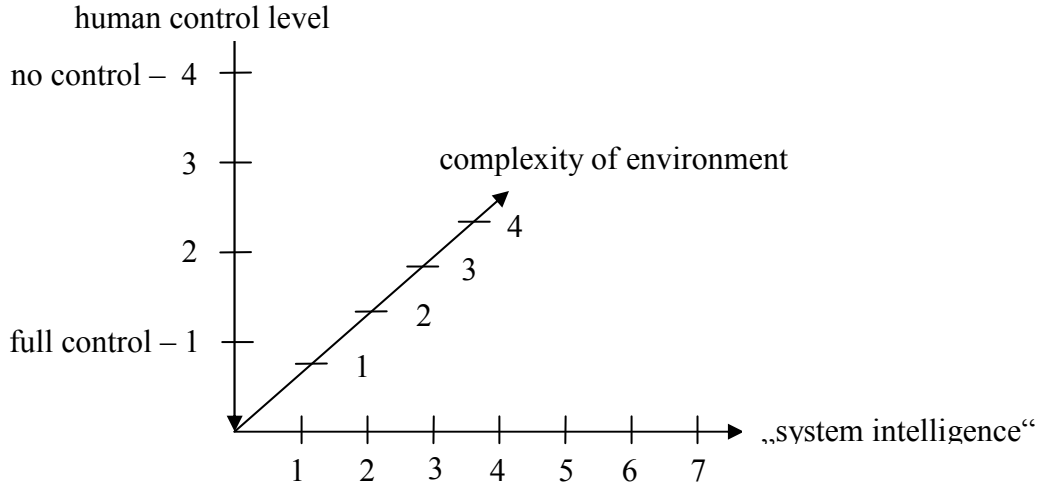


Figure 2: Schematic diagram of the three-dimensional vector space.

By allocating existing or future automated weapon systems within the three dimensional space, it will be possible to distinguish systems by their overall level of automation/ autonomy. Within this approach, the scaling of the three vectors (see table 2) should rather be seen as a first attempt of classification, which indeed has to be further discussed and enhanced over time.

Level	***… of "system intelligence"***
1	very simple (e.g. binary sensing)
2	able to differentiate and assess simple static situations
3	able to differentiate and assess more complex static situations
4	able to differentiate and assess simple dynamic situations
5	able to differentiate and assess more complex dynamic situations
6	able to differentiate and assess a broad range of complex dynamic situations (by its programming)
7	able to differentiate and assess all complex dynamic situations like a human operator (by self learning)
Level	***… of intended human control*** *(decreasing)*
1	total human control over all system actions (remote control or hard-coded actions in advance)
2	partial human control (e.g. control over target acquisition and/or weapon engagement)
3	human supervision over system actions (ability to override or abort system actions)
4	no human control over time and nature of system actions
Level	***… of complexity of the operational environment***
1	static operation on battlefield (having a fixed perimeter)
2	mobile operation on battlefield
3	operation in a contested environment
4	operation in a civil environment

Table 2: Exemplary scaling of the three vectors.

In order to give an impression on how to apply the three-dimensional approach by allocating and distinguishing weapon systems, four exemplary systems are chosen and – dependent on the individual assumed characteristics of each weapon system – are assigned with a certain value for each of the three independent vectors spanning the three-dimensional space (see table 3). The chosen values for each system were estimated and have to be seen exemplarily. Next, the figures 3 and 4 show the allocation of the exemplary weapon systems within the space and illustrate their spatial discrimination.

System	Complexity of the operational environment	Intended human control	System intelligence
MQ-9 Reaper	2/**3**/4	2	4
Anti-personnel mine	1	4	1
Close-In Weapon System (CIWS), e.g. Phalanx	1	3	4
Cruise Missile	2/3/**4**	1	3

Table 3: Exemplary vector values for different weapon systems depending on assumed characteristics and mission tasks.

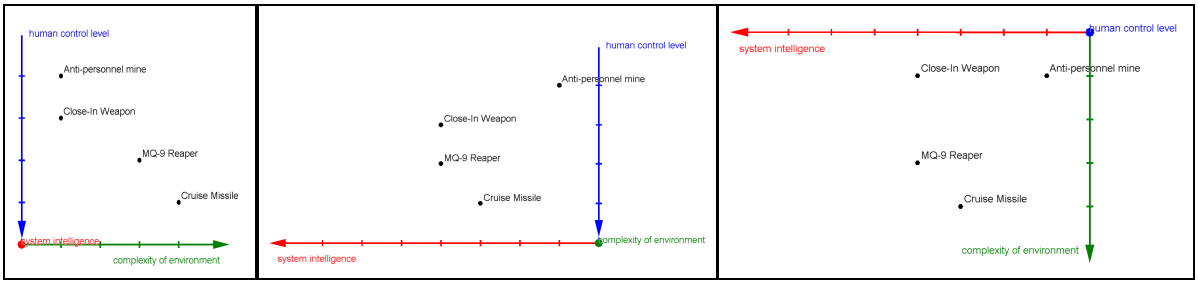


Figure 3: Location of the exemplary weapon systems, in each case related to only two axis of the three-dimensional vector space.

Furthermore, thereby it could be possible to identify domains within the three-dimensional space where the use of a system may be problematic concerning norms and laws (e.g use of a dump system acting on its own in a complex civil environment). So, depending on its level of automation the individual system's compliance with humanitarian and international law might be evaluated for each area of intended use respectively operational environment this way.

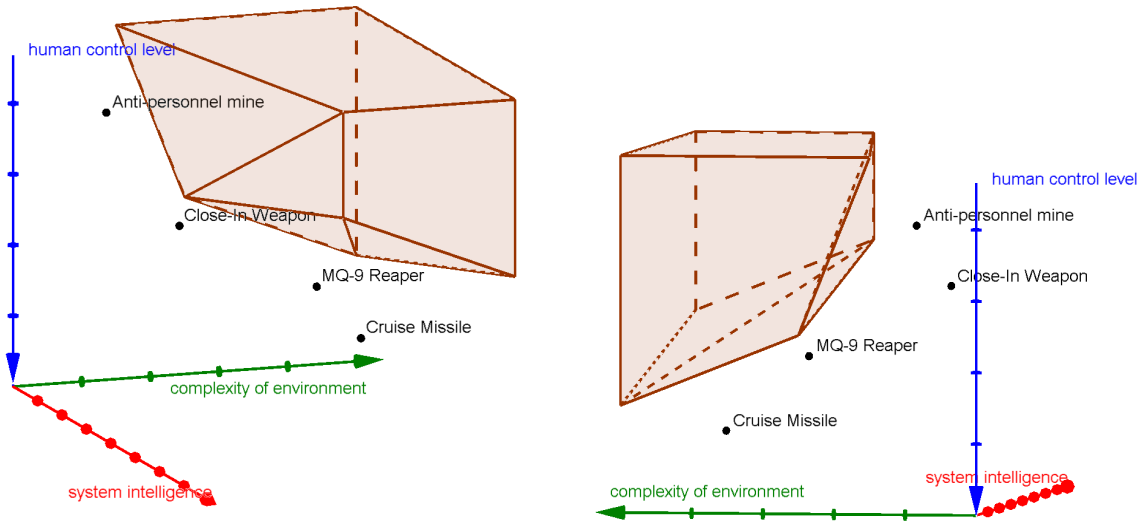


Figure 4: Overall 3D image of the three-dimensional vector space and the individual located weapon systems. Furthermore a hypothetical domain (geometric figure) where the system use may be problematic with regard to norms and laws.

6. Summary

Both approaches – intended to classify weapon systems regarding their level of automation – shall open up new perspectives on the term *autonomy* and thus feed the ongoing discussions related to LAWS. But the authors of this paper do not claim to exclusively answer the question how *autonomy is defined*. A crystal clear distinction between *automatic* and *autonomous weapon systems* is not provided by both approaches, which seems to be very challenging or even impossible due to the very nature of the problem. Different levels of *automation* are rather defined by complex software configurations and source code than specific hardware components. At the same time, this makes it very difficult to supervise and control armaments efforts in this field, as demonstrated at the CCW meetings. The potential risks related to *autonomous killer robots* require an intensification of efforts to come to a better understanding of the term *autonomy* when applied to *weapon systems*, which constitutes a crucial pre-requisite for further discussions on LAWS and an agreement regulating their development and fielding preventively on an international level.

IFSH, February 2016

Further information: http://www.ifsh.de/IFAR

Contact:

Dr. Christian Alwardt	email: alwardt@ifsh.de	Tel. +49 (0)40 866077 - 77
LtCol (GS) Martin Krüger	email: krueger@ifsh.de	Tel. +49 (0)40 866077 - 54