University of Zurich^UZH

UZH Digital Society Initiative

An Evaluation Schema for the Ethical Use of Autonomous Robotic Systems in Security Applications

Markus Christen
Thomas Burri
Joseph Chapa
Raphael Salvi
Filippo Santoni de Sio
John Sullins

DSI White Paper Series

White Paper No. 1

Table of Contents

Introduction

Information technology has become a decisive element in modern warfare, in particular when armed forces of developed countries are involved. Modern weapon systems would not function without sophisticated computing power, but also the planning and executing of military operations in general heavily rely on information technology. In addition, armed forces, but also police, border control and civil protection organizations increasingly rely on robotic systems with growing autonomous capacities. This poses tactical and strategic, but also ethical and legal issues that are of particular relevance when procurement organizations are evaluating such systems for security applications.

In order to support the evaluation of such systems from an ethical perspective, this report presents an evaluation schema for the ethical use of autonomous robotic systems in security applications, which also considers legal aspects to some degree. The focus is on two types of applications: First, systems whose purpose is not to destroy objects or to harm people (e.g. rescue robots, surveillance systems); although weaponization cannot be excluded. Second, systems that deliberately possess the capacity to harm people or destroy objects – both defensive and offensive, lethal and non-lethal systems. The cyber-domain where autonomous systems also are increasingly used (software agents, specific types of cyber weapons etc.) has been excluded from this analysis.

The research that has resulted in this report outlines the most important evaluations and scientific publications that are contributing to the international debate on the regulation of autonomous systems in the security context, in particular in the case of so-called lethal autonomous weapons systems (LAWS). The goal of the research is twofold: First, it should support the procurement of security/defense systems, e.g. to avoid reputation risks or costly assessments for systems that are ethically problematic and entail political risks. Second, the research should contribute to the international discussion on the use of autonomous systems in the security context (e.g., with respect to the United Nation Convention on Certain Conventional Weapons). In this way, the report should meet the information needs of armasuisse Science + Technology and related institutions of the Swiss government such as the Arms Control section of the Swiss Department of the Exterior and the Arms Control and Disarmament section of the Federal Department of Defense.

This report results from a research project funded by armasuisse Science + Technology, the center of technology of the Federal Department of Defence, Civil Protection and Sports. The research was conducted by a team of the Center for Ethics of the University of Zürich (principal investigator PD Dr. Markus Christen; research assistant: Raphael Salvi) and with the support of an international expert team. This team consisted of Prof. Thomas Burri (University of St. Gallen; focus on chapter 3 of part 2), Major Joe Chapa (United States Air Force Academy, Department of Philosophy; focus on chapter 2), Dr. Filippo Santoni de Sio (Delft University of Technology, Department Ethics/Philosophy of Technology; focus on chapters 1 and 4), and Prof. John Sullins (Sonoma State University, Department of Philosophy; focus on chapter 4). The report was reviewed and corrected by the whole team.

The research relied on an extensive literature search based on the knowledge of the expert team, on 21 interviews with external experts (technology, law, military, ethics), and on the feedback obtained during a two-day workshop in Zürich. The workshop included internationally renowned experts in the field and agents from interested entities of the Swiss government and the International Committee of the Red Cross (ICRC). The involvement of these external persons in the workshop does not indicate their approval – or the approval of the entities they represent – of the content of this report. They were only consulted as external experts and were not asked to endorse the finding of this report.

The report is structured as follows. The **first part** of the report outlines the proposed evaluation schema that consists of three steps: deciding about the applicability of the evaluation schema (step 1), deciding about the use intention of the robotic system (step 2), and – depending on step 2 – the analysis of the system under consideration based on the criteria. The **second part** provides background information regarding the evaluation schema. The *technology chapter* focuses on relevant technologies used in autonomous systems, degrees of system autonomy and likely developments of applications in the security domain in the next 10-15 years. The *security chapter* discusses types of autonomous systems in the security sector as well as the status of autonomous capacities in military command and control structures, including an outlook on future developments. The *law chapter* focuses on the current international debate on regulating autonomous systems, explains the main legal issues autonomous weapons systems raise, and briefly discusses possible developments. The *ethics chapter* of the report outlines the current ethical discussion on system autonomy, discusses major pro- and con-arguments and sketches likely developments. The *materials chapter* lists the persons interviewed, the workshop participants and the literature used.

Disclaimer: The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Federal Department of Defence, Civil Protection and Sport.

Please cite as: Christen, Markus; Burri, Thomas; Chapa, Joseph; Salvi, Raphael; Santoni de Sio, Filippo; Sullins, John (2017): An Evaluation Schema for the Ethical Use of Autonomous Robotic Systems in Security Applications, November 1 2017, UZH Digital Society Initiative White Paper Series No. 1, University of Zurich. Available at SSRN: https://ssrn.com/abstract=3063617

Feedback welcome: The author team appreciates any feedback that helps to improve this report. Please send comments, corrections and suggestions to the Principal Investigator of this study: Markus Christen, christen@ethik.uzh.ch

Part 1 – Evaluation Schema

Preliminary remark: The purpose of this schema is to help identify ethical issues that the use of autonomous robotic systems can give rise to when they are deployed in defined circumstances within the security sector. The evaluation schema intends to inform the procurement process of such systems. The evaluation schema is not a decision algorithm of which the output would determine whether a system is ethically problematic or unproblematic. Rather, it points to issues that require further analysis in the assessment of autonomous robotic systems in security applications. Furthermore, ethical issues will have to be balanced with other relevant aspects for the decision process such as financial, legal or technological aspects.

Part 2 of this report comprises detailed background information to the various issues that are addressed in the evaluation schema. Grey boxes indicate references to sections in part 2 of this report, where the reader can find additional information.

General Outline of the Evaluation Schema

Before proceeding to apply the evaluation schema, four aspects need to be highlighted:

1) First, one has to evaluate whether the system under consideration has (minimal) **capacities usually attributed to robotic systems** and a **sufficient degree of autonomy** in order to fall into the domain of application of this evaluation schema. Tools or weapons that are under complete human control or only perform simple automated procedures are not the concern of this evaluation schema, although they certainly can raise ethical or legal issues. The degree of autonomy is assessed in this evaluation schema along the following criteria:

   - **Autarchy:** The robotic system has some degree of autarchy with respect to energy supply or other resources that are essential for its functioning.
   - **Independence from human control:** At least some functions of the robotic system are performed without any human intervention (e.g., gait in a walking robot), although higher-level of control is still possible.
   - **Interaction with environment:** The robotic system is equipped with sensors and effectors that allow for some interaction with a (changing) environment, objects, humans, or other robotic systems. This may include defense abilities against hostile behavior.
   - **Learning:** The system is equipped with some capacity to learn from data provided by external sources, or by data that the system itself is recording.
   - **Mobility:** The robotic system is able to move in a (defined or restricted) geographic area of a certain complexity and for a certain time.

Those criteria for assessing autonomy of robotic systems are derived from a larger set of dimensions that are discussed within the technical literature. An introduction into the main topic of robotics is provided in part 2, **chapter 1** of this report; section 1.3 provides a discussion of dimensions of system autonomy.

2) There are **two classes of evaluation criteria**. Which class is applied depends on the intention for which the robotic system under consideration has been designed. In this way, the evaluation schema takes into account that it matters, from an ethical point of view, whether a robotic system deliberately includes capacities to harm people or to destroy objects (i.e., systems that are **weaponized**[1]), or whether the possibility that a robotic system could harm or destroy is an unwanted side-effect of its deployment. Thus, the first step in applying this evaluation schema is to decide into which of the two categories the robotic system falls.

   - If the robotic system **is not intentionally designed to include capacities to perform operations directly aimed at harming people or destroying objects**, a first criteria set A comes into play. This set of criteria takes into account that any real-world robotic system that interacts with its environment could harm people or destroy objects, either due to malfunction or due to unexpected circumstances for which the system was not designed. As the security context generically involves situations (e.g., rescue missions, supply missions in combat, etc.) where the likelihood of severe ethical consequences is higher than in other contexts (e.g., robotics applications in manufacturing), this evaluation schema realistically factors in the risks involved when deploying those systems and the potential for dual-use; i.e. the likelihood that the robotic system can be redesigned into a system such that criteria set B (below) would come into play.
   - If the robotic system **is intentionally designed to include capacities to harm people or destroy objects**, then a criteria set B comes into play <u>in addition to set A</u>; i.e. those systems should be evaluated with respect to both the A and B criteria. This set of criteria takes into account that systems deployed with harmful (or even lethal) capacities are generally used in situations of highest ethical concern and require a more sophisticated evaluation by law[2]. Criteria set B takes into account the capacity of a system to comply with ethical requirements that are in line with accepted ethical norms such as human rights.

Generally, this report covers applications of autonomous robotic systems in the security sector – a topic outlined in part 2, **chapter 2**. A detailed definition of "security sector" is provided in section 2.1. Examples of current autonomous systems used in the security sector are given in sections 2.2 and 2.3.

3) The criteria applied in this system are **not equally determinable**. This results from the fact that the legal norms and ethical principles that inhere in these criteria are usually formulated on an abstract level and they are not in all cases sensitive to differences in context. This means that the evaluation schema includes estimations of how credibly and reliably each criterion can be applied by its users. Five different groups of criteria will be distinguished, although some overlap between those groups can be expected:

   - **Criteria related to the physical characteristics of a robotic system:** These criteria are expected to be relatively easy to apply and to lead to credible and reliable results. For example, they refer to the presence of certain physical safeguards to prevent accidents or to design aspects that prevent certain types of misuse.

[1] According to the Oxford Dictionary, a weapon is a thing designed or used for inflicting bodily harm or physical damage.

[2] Article 36 of the 1977 Additional Protocol I of the Geneva Conventions requires states to review new weapons, means and methods of warfare.

- **Criteria related to the behavioral characteristics of a robotic system:** These criteria refer to the interaction of the robotic system with its environment, persons other than the system operators, objects or other robotic systems. They include an evaluation of the software that controls the system and simulation possibilities in order to assess system behavior. It is expected that these criteria are more difficult to determine, in particular when the software involves some learning capacity.
- **Criteria related to the operator of a robotic system:** These criteria refer to control possibilities, human factor issues, possible training of the system and the associated training requirements for the operators. We expect that those criteria are comparably easy to determine as they refer to standard conditions systems have to meet when being available on a market.
- **Criteria related to the deployment conditions of the robotic system:** These criteria refer to the context in which the system is planned to be used and to the possibilities to constrain the system activity with respect to geographical, temporal or other factors. Given the uncertainty related to the use of autonomous systems and the potentially high variety of contexts, we expect these criteria to be more difficult to be determined.
- **Other criteria:** Some additional criteria are not covered by this classification but they still are relevant for robotic systems. Examples include the data the systems generate which may involve data protection issues or non-proliferation issues (i.e., preventing an increase of countries possessing autonomous weapons).

The criteria used refer mainly to ethical considerations; an introduction into the ethics of autonomous systems is provided in part 2, **chapter 4** of this report.

4) The application of the evaluation schema results in **three evaluation outcomes** for each criterion, based on an (extended) traffic light rating (also called "red-amber-green" or "RAG" rating). RAG rating is a widely used and easily understandable way for indicating the status of a variable with respect to danger, performance etc. In this schema, we use an adapted RAG rating that includes "grey" to denote the fact that a criterion is not applicable in a certain case. The RAG ratings applied in this report yields the following:

- **Green:** This rating results when the system fulfills the criterion with a sufficient degree of reliability, taking the difficulty of measuring the criterion into account. Difficulties to measure criteria are mirrored in a "best practice" approach. For example, if the system behavior is assessed using a simulation approach, the type of technology used to perform the simulation is likely to change (and improve) in time. "Best practice" thus means that the currently available best approach for simulating system behavior is used, leading to a "green" evaluation if the test is passed successfully. Future simulation methods, however, could lead to a different result. Thus, a "green" rating should not be understood as a perennially valid outcome.
- **Amber:** This rating applies when a) there is considerable doubt that the system fulfills the criterion, or when b) the uncertainty whether a criterion is fulfilled is too high to allow for a credible rating.
- **Red:** This rating applies when the system fails to comply with the criterion or complies only with an insufficient degree of reliability. Again, this rating is not perennial. If later the

technology used to measure the criterion turns out to have some flaws, a re-assessment becomes necessary.

- **Grey:** This rating applies when a criterion is inadequate to be used for a certain robotic system. Criteria rated as "grey" are not considered for the overall assessment of the system.

Depending on the specific case, some criteria may be more relevant than other criteria, leading to a weighing of the criteria (high, medium, or low). Applying the evaluation schema thus leads to a set of green, amber and red ratings that allow for an overall assessment of the robotic system. ***This overall assessment is not intended to make a clear statement that a system is ethically acceptable or not in the sense that exceeding a certain threshold for the number of "red" ratings generates a "no go" statement. Rather, the more amber or red ratings are generated during an evaluation, the higher is the need for justification if one still wants to deploy the system. Providing the justification is not the aim of this schema.***

In contrast to this evaluation schema, a legal analysis is required to yield clear statements regarding the acceptability of a weaponized autonomous robotic system. According to Article 36 of the Additional Protocol I to the Geneva Conventions, each State Party is required to determine whether the employment of a new weapon, mean or method of warfare that it studies, develops, acquires or adopts would, in some or all circumstances, be prohibited by international law. This evaluation schema is not intended to replace standard weapon review processes within this legal framework, but rather to supplement them and highlight ethical concerns. It is embedded in the current discussion within international humanitarian law outlined in part 2, **chapter 3** of this report.

Step 1: Deciding about the Applicability of the Evaluation Schema

The first step is to decide whether the robotic system under consideration falls into the scope of this evaluation schema. First, this requires that the system can (in a reasonable sense) be called a "robotic system". Such a system is expected to possess – at least to a minimal degree – the following capacities:

- **Sensing:** The system receives sensory input allowing it to gather some information emerging from the environment of the system.
- **Computing:** The system is equipped with certain algorithms and software in order to control the behavior of the system.
- **Effecting:** The system has some capacities to influence its environment physically through effectors.
- **Communication:** The system has some capacities to communicate (i.e., accept orders or inform about its inner state) with humans or other systems.

Systems that lack one of these capacities do not fall into the scope of the evaluation schema. Rudimentary capacities are sufficient, though.

More information on the definition of robots and robotic systems and enabling technologies of robotic systems is provided in part 2, Chapter 1, sections 1.1 and 1.2 of this report.

Second, the robotic system needs to possess a **certain degree of autonomy**. The notion of system autonomy is widely debated in the robotics community and beyond. This first step assesses the autonomous capacity of the system along the following five dimensions that condense this discussion into properties that are relatively easy to evaluate. Here, the purpose is not to measure the ethical impact a system may have. Rather, the dimensions help the evaluator acquire a sense of the degree of autonomy of a system.

	Degree of fulfillment of this dimension is…		
	…low	**…medium**	**…high**
Autarchy *This criterion concerns the capacity of the system to function independently from external energy sources.*	The system does not include any built-in capacities to replace energy needed to function, after standard resources (e.g., fuel in the tank) have been exhausted and it completely depends on supply from external parties.	The system has some internal fallback options to access resources it needs for performing its task and it can access these resources in dependence from externally changed circumstances.	The system has built-in capacities to retrieve or replace energy resources if needed (e.g., solar cells) and it can actively seek resources it needs for performing its function.
Independence from human control *This criterion concerns the degree upon which the functioning of the system depends on human action or intervention.*	The system's functions and activities are under the complete control of a human operator, except for simple automated responses.	The system performs some of its sub-routines independently from a human controller. It may operate in physical distance from the operator, although the operator has access to the main performance of the system and is able to intervene most of the time.	The system can conduct a substantial part of its operational duration without human interference and in physical distance from the operator. In case of unforeseen circumstances, the system is able to request help from the operator or to rely on fallback options (e.g., return to base).
Interaction with environment *This criterion concerns various types of interaction of the system with its environment.*	Sensors and/or effectors of the system only serve simple signaling or simple automated responses. Defensive means are purely passive (e.g., passive armor).	The system is equipped with capabilities that allow for an interaction within a structured environment. It includes defensive means that can adapt to some degree to the external environment.	The system is equipped with sensors and effectors that allow for an interaction with an unstructured environment. It has a sophisticated repertoire of defensive means that can be used flexibly.
Learning *This criterion concerns the capacity of the system to adapt its programming and behavior based on the previous data acquired.*	The system's behavior is completely determined by internal programs or human commands. The system is unable to learn from past interactions.	The system has some capacity of learning and is able to adapt its behavior based on previous experience of itself or others. The adaptations are reasonably comprehensible for humans in charge, or the training process is suspended and the system is tested/evaluated before fielding.	The system is equipped with sophisticated machine learning capacities, allowing it to actively perform some interaction in order to explore the environment and to learn from experience. The learning of the system and the resulting operation is explainable only with great effort over a long period.
Mobility *This criterion concerns the capacity of the system to displace itself.*	The system is immobile unless transported by external means.	The system is able to move in restricted, pre-defined, structured environments.	The system is able to move in a variety of different environments that involve some degree of contingency.

Robotic systems that score "low" in all five dimensions do not fall into the scope of this evaluation schema, whereas robotic systems that score "high" in at least one dimension necessarily fall into the scope of the evaluation schema. In-between cases are evaluated on a case-by-case basis. In cases of doubt, we recommend the application of the evaluation schema.

More information on the concept of autonomy in robotics is provided in part 2, section 1.3 of this report.

Deciding whether a system falls into the category of an "autonomous robotic system" is not a simple task, because autonomous capacities have a long history in the security sector. For example, anti-personnel mines are sometimes regarded as rudimentary autonomous weapons. However, because they are simply designed to explode based on the presence, proximity, or contact of a person or vehicle, which is a simple automated response, and the relevant decision is one of the operator – namely where to put the mine, the mine would score low on the dimension "independence from human control" (furthermore, a mine is not a robot, i.e. would fall out of the evaluation scheme anyway). Other weapons may have more sophisticated autonomous capacities. For example, they may be designed so that they are able to select or engage targets automatically after having been activated by the user. The United States (and other countries) has at its disposal weapon systems for local defense with autonomous capabilities designed to counter time-critical or saturation attacks. These weapon systems include the Aegis ship defense system and the Counter-Rocket, Artillery, and Mortar (C-RAM) system (United States Department of Defense 2015; section 6.5.9.1.).

While standard mines do not fall into the scope of this evaluation schema, it is clear that mines as well as other weapon systems that lack autonomous capacities as described above can pose serious ethical problems. In other words, a system that is excluded from the scope in this step is not necessarily "ethically safe". Furthermore, some of the criteria applicable in step 3 below may also be relevant when assessing systems that lack autonomous capacities. Nevertheless, as the aim of the evaluation schema is to evaluate *autonomous robotic* systems, the evaluator should first get a sense whether the robotic system in question is indeed an autonomous robot to a degree that warrants an evaluation.

More information and detailed examples on recent developments in robotic systems in general are provided in part 2, section 1.4. Examples referring to autonomous systems in the security sector are provided in part 2, sections 2.2 and 2.3 in part 2 of this report.

Step 2: Deciding on the Design and Use Intention of the Robotic System

The second step is to decide whether the robotic system under consideration was designed with the intention to harm persons or destroy objects. Although the main intention embodied in the design of the system need not be to harm, the spectrum of actions the robotic system can perform may be such that the intention to harm is included. The intention to harm, in other words, may be subordinate and depend on circumstances. An example would be a guard robot. The main intention for deployment of this robot may be to protect a certain building, but it may have the capacity to deploy force after an intruder fails to react to several warning messages.

However, we can expect the following "grey areas" with respect to a system's design and intended uses:

- Some robotic systems may be equipped with effectors that induce *psychological* harm in people, e.g. by blinding people with a bright light, alarming a person using acoustic means (e.g., a loud whistle) or by discharging olfactory substances. Also, the appearance matters: Some robots are cute and cuddly and some look like the "Terminator" – the latter likely express the intention to create psychological harm. While the use intention is most relevant with systems equipped with such effectors, they should be considered as designed to harm, so long as there is a foreseeable pathway between the intended use of the effector and the harm, depending on the specific weapon (eye injury in case of lasers, ear damage in case of acoustic noise, irritant effects to skin/eyes from olfactory substances). A system is not harmful for the purpose of this step, merely because a person experiences a psychological shock upon encountering it (if the system is otherwise harmless). Such a system lacks effectors, but a foreseeable pathway between use and harm is also absent – unless the system was designed to shock by appearance.

- Some robotic systems may be equipped with purely defensive functions against aggressors that threaten the integrity of the system. For example, the system could activate a shield against physical impact, involving the possibility that the attacker is harmed when the shield is extended. Depending on the aggressiveness of defense (e.g., how fast the shield is unfolding), a system may be considered harmful for the purpose of the present step.
- Some robotic systems may have such physical properties that there is an inherent risk of injury, death or destruction (e.g., a person might die in a crash; the system might roll over a person, etc.). It may also be impossible or reasonably unfeasible to secure systems against accidents or hacking. However, the more the harm is foreseeable and reasonable measures to prevent it are ignored, the more the systems may be considered as intended to harm for the purpose of the present step.

Any robot has the potential to induce psychological harm in persons interacting with the system, depending among other things on the psychological vulnerability of the persons involved. Even very simple protective operations of a robot could in some cases harm people. Every robot of a certain size is capable of physically harming people, especially when it is malfunctioning or being misused. Given this, the presence of protective operations or size alone should not qualify a system as being "intended to harm". Such aspects will nevertheless be relevant when it comes to assessing systems not intended to harm.

The following points should help evaluate whether a robotic system qualifies as a system designed with the intention to harm (or used with this intention, if the system is equipped with those devices by the user after procurement). The points relate to whether the system is equipped with some type of weapons; i.e. with something designed or used for inflicting bodily harm or physical damage, which is a clear sign that an intention to harm is present. In cases in which it is not clear how to qualify a robot's properties (e.g., a threatening looking robot's warning tone may sometimes, but not always terrify a person), the intent to harm should be considered "Unclear".

In this step "misuse" or "dual use" of a robot should not be factored in, unless reasonably foreseeable. For example, when a bomb is transported by a rescue robot or an autonomous truck has been hacked to overrun pedestrians, this would not qualify as a "yes".

Presence of capabilities to harm:	**No**	**Unclear**	**Yes**
The robotic system is equipped with a kinetic weapon (gun, rocket launcher, explosives, etc.) or is designed in a way that such a kinetic weapon can effortlessly be integrated into the system architecture.			
The robotic system is equipped with an effector that targets the sensory nervous system or central nervous system of humans and that is able to create temporary or permanent damage on the human sensory system (e.g., blinding laser, very loud acoustic stimuli; hazardous gas; paralysis inducing tool, etc.).			
The robotic system is designed in a way that it explicitly intends to terrify human beings and to bring them into a state of psychological stress, trauma, and the like.			
The robotic system is equipped with an effector that has an otherwise destructive effect on humans or objects (e.g., biological or chemical agents, microwaves, EMP generator, high-energy-laser, etc.).			

Only if all of these questions yield a "no", the criteria system A is sufficient for evaluating the system, otherwise criteria system A and B applies.

Whether a robot should qualify as "intended to harm" also depends on the context in which it is likely to be deployed. Generally, in any civilian setting (police, border control, disaster management), it is less likely that autonomous robotic systems will be used with the intention to harm, whereas in a military setting, this is much more likely. The different legal rules governing the two domains are a testimony to this distinction (intentional harm is more strictly prohibited in civilian settings). Whether a robot should qualify as "intended to harm" may thus vary depending on the foreseeable use in different contexts. The use in a certain context may make certain intended uses more likely and this may reflect back on the characterization of the robot as intended to harm. Caution therefore needs to be applied when a system is moved from one context to another, e.g. when a robotic system is decommissioned from the military with a view to be used by police (or the other way around). The full evaluation schema then needs to be applied again.

More information on the different legal rules applying to autonomous robotic systems is provided in part 2, **chapter 3** of this report, specifically in section 3.2.

Step 3-A: Applying the Criteria for Robotic Systems Not Intended to Harm

The following criteria set A comes into play when a robotic system is not intentionally designed to include capacities to perform operations directly aimed at harming people or destroying object.

The criteria in Step 3-A apply to <u>all systems</u>, independent from the fact that they are designed with the intention to harm or not.

Systems that are designed with the intention to harm also have to be checked with <u>additional criteria</u> (Step 3-B).

The following points are important when applying the evaluation schema:

- The evaluation schema does not include a final, exhaustive list of criteria. Depending on technological progress, additional criteria may be needed, whereas some criteria may lose importance.
- Some of the criteria may also be relevant for the assessment of non-autonomous robotic systems that are not included in the scope of this evaluation schema.
- A single "red" label does not imply that the use of a system is necessarily ethically impermissible. The number of green, amber and red labels instead provides an indication of the justificatory pressure with respect to the ethical use of a system.

For each criterion, only very basic information regarding ethical importance is provided here. Detailed background information on ethical aspects is provided in part 2, chapter 4 of this report.

Physical Characteristics of the Robotics System

Appearance of the robotic system

Core Question: To what extent is the physical appearance of the robot (e.g. shape, color) likely to trigger only appropriate (as opposed to hazardous and undesired) emotional and behavioral reactions in the human user/interacting person?

Determination: Visual cues are a key component of industrial design. They concern various facets such as color, size, shape, texture, etc. Those physical elements of the appearance of the system should properly trigger the appropriate psychological and emotional states in humans that interact with the system, so as to prevent hazardous and other undesired behavior by the interacting persons. This can be determined with user surveys or more elaborated testing.

Weight: The more the system is expected to interact with non-expert and/or vulnerable users (e.g. persons under stress, children, and elderly persons) the more relevant is this criterion.

Green: The interaction of the system with humans has been systematically studied and tested with a wide range of users in all the potential contexts of use and no relevant inappropriate, hazardous or undesired emotional or behavioral response associated to the physical appearance of the system has been observed or has been predicted to occur.	**Amber:** The interaction of the system with humans has been studied and tested but not systematically and/or not in all the potential contexts of use and/or a limited number of inappropriate, hazardous or undesired emotional or behavioral responses associated to the physical appearance of the system have been observed or have been predicted to occur.	**Red:** The interaction of the system with humans has been hardly or not at all studied and tested and/or in the studies and tests some inappropriate, hazardous or undesired emotional or behavioral responses associated to the physical appearance of the system have been repeatedly observed or the risk of their occurrence has been deemed high.	**Grey:** The system is expected to interact only with highly trained specialized personnel.

Physical safeguards

Core questions: To what extent do physical safeguards exist, that ensure that the operator(s) of the system or persons that likely will be exposed to the robot cannot interfere with mechanical parts of the robot (e.g., rotor protection)? Alternatively, if they can, do such safeguards provide sufficient warning from potential dangers?

Determination: Physical safety is a standard requirement for robotic systems and all aspects of physical safety should be adequately described in the user manual.

Weight: The more moving parts a robot has and the more kinetic energy the movements of those parts involve, the more relevant is this criterion.

Green: Physical safeguards are provided that are adequate for the functional properties of the system.	**Amber:** There is some apparent lack of physical safeguards, but the risk of causing harm is small.	**Red:** There are clear risks that the operator(s) can be harmed by the robot due to the absence of physical safeguards.	**Grey:** The robot has no relevant mechanic parts that could hurt a human or the maximal kinetic energy is too small to generate damage.

Behavioral Characteristics of the Robotics System

Autarchy

Core Questions: Does the system operate in a largely autarkic manner? Does it re-supply energy from sources that are not subject to human control?

Determination: This requires examination of the systems energy supply (battery, tether, fuel, etc.) and the way it is re-charged (if applicable); determination of time period of self-sufficiency (possibly under varying circumstances).

Weight: The longer a system is capable of operating without human feedback/intervention, the more important the criterion.

Green: System is not in any way self-sufficient; energy supply can be cut physically at any time.	**Amber:** System is capable of operating autarkically for some limited, clearly determined time,	**Red:** System is autarkic for long periods ("loitering"); human intervention is impossible for	**Grey:** System is purely mechanical. (Note that it would then not come within the scope of the evaluation.)

<table>
<tr><td></td><td>during which human intervention is always possible.</td><td>longer than just very brief intervals (e.g. underwater systems)</td><td></td></tr>
<tr><th colspan="4">Behavior recorder</th></tr>
<tr><td colspan="4"><b>Core Questions:</b> Is an electronic recording device available in the robot that stores data on the major behavioral activities of the robot? In case of incidences, does this data allow to reconstruct the event and help to identify responsibilities? Are the access rights to this data determined (e.g., to legal entities in case of accidents)?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Check which variables the behavior recorder is storing and evaluate, whether those data indeed determine the behavior of the robot or whether emergent behavior can emerge that is not captured by the data. This includes testing in possible accident situations and reconstruction of the accidents based on the data. Check whether the behavior recorder is sufficiently secured against physical damage or data manipulation (e.g., through encryption). Check the data management plan of the behavior recorder with respect to data capacity, long-term data storage and access of data (by whom, etc.).</td></tr>
<tr><td colspan="4"><b>Weight:</b> The higher the liability risks and the more likely it is that the system operates in an environment where incidences of high ethical risks can happen, the higher is the weight of this criterion.</td></tr>
<tr><td><b>Green:</b> The behavior recorder stores the relevant data in a safe and secure way; the data management plan involves all relevant cases.</td><td><b>Amber:</b> There are questions about whether the behavior recorder stores the relevant data; there are privacy and/or security risks.</td><td><b>Red:</b> No behavior recorder is available or the behavior recorder is insufficiently secured.</td><td><b>Grey:</b> The type of constraints under which the robot operates is incompatible with including a behavior recorder.</td></tr>
<tr><th colspan="4">Deception</th></tr>
<tr><td colspan="4"><b>Core Question:</b> If the robotic system has been designed for affective and emotional interaction with the user and other agents who may interact with it (for instance in a police and rescue operation): Is the degree of deception involved controlled and justified?</td></tr>
<tr><td colspan="4"><b>Determination:</b> This requires first a theoretical evaluation of what kind of deception is possible and warranted in the application context of the system. Deception has to be distinguished from general questions regarding the psychological impact of the system, as deception is an intended effect; i.e. one wants that the interaction partner has some beliefs with respect to the system that the system actually does not fulfill. Therefore, one has to answer three questions: First, did the designer intent to deceive the interaction partner (requires inquiring of the producer/designer)? Second, does deception actually work as intended (requires experimental studies)? Third, is deception ethically warranted in this situation (requires a theoretical/legal analysis)?</td></tr>
<tr><td colspan="4"><b>Weight:</b> The weight of this criterion depends on several aspects: First, does the context allow for some degree of deception (e.g., a police operation involving a suspect, level of emergency). Second, can we expect an implicit consent for being decepted? Third, how vulnerable is the intended interaction partner?</td></tr>
<tr><td><b>Green:</b> Interaction design has been tested and possible deception is ethically justified.</td><td><b>Amber:</b> Insufficient testing of interaction design; open questions regarding deception.</td><td><b>Red:</b> Unjustified deception.</td><td><b>Grey:</b> The robot has not been designed for emotional interaction.</td></tr>
<tr><th colspan="4">Dilemma behavior</th></tr>
<tr><td colspan="4"><b>Core Questions:</b> Will the system operate under conditions where ethical dilemmas may occur; i.e. decision situations, where any option, even inaction, will likely cause some harm (for instance, deciding which areas to explore first in a rescue operation among two or more that are affected by a disaster)? Does the robotic system have built-in options/procedures or triage protocol in order to decide when being confronted with a dilemmatic situation? Are those procedures ethically justified?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Simulation of system behavior under conditions that involve dilemmas. Analysis of built-in decision procedures.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The more often dilemmas can be expected and the more impact the decisions have, the more relevant is the criterion. If the robot has built in procedures etc. to take potentially harming decisions in complex scenarios, then this criteria becomes more relevant and the results of those decisions need to be tested.</td></tr>
<tr><td><b>Green:</b> The robotic system is to some degree able to predict the likelihood of dilemmatic situations and can inform operators for guidance in advance. If the system has to react autonomously, it makes decisions that</td><td><b>Amber:</b> The system is unable to cope with dilemmas and stops or withdraws its operation completely. This may be a problem as doing nothing can be worse than choosing an imperfect solution (e.g., leaving both victims to die if not being able to choose</td><td><b>Red:</b> The robotic system systematically makes decisions that are inconsistent with the procedures or protocols that a professional trained human would follow in a similar context or that can’t reasonably be justified or comprehended. And/or: there is no way</td><td><b>Grey:</b> The robotic system is not operating under conditions where one reasonably can expect dilemmas.</td></tr>
</table>

informed humans can comprehend or can reasonably be justified. The results of the actions are consistent with the procedures/protocols that a professional trained human would follow in a similar context.	which one to save). Putting back the operator in charge (if possible) does not allow for improving the handling of the dilemma.	to predict how the system will behave in dilemmatic circumstances. And/or: the decision-making of the system is not sufficiently transparent (e.g., due to the learning mode applied by the system).	

General safety feature testing

Core Questions: Has an initial operational test and evaluation been performed upon delivery of the system to ensure that critical safety features work as intended? Does the supplier provide methods to regularly test the software prior to a mission to validate that critical safety features have not been degraded?

Determination: Simulation respectively output of certified Testing &Evaluation routines provided by the manufacturer.

Weight: The more the system is operating in an environment with potentially high collateral damage, the higher is the weight of this criterion.

Green: Initial operational test and evaluation has been performed. Integrated testing routines are available for all critical safety features.	**Amber:** Testing and evaluation is performed and routines are provided, but not all critical safety features are covered.	**Red:** No testing and evaluation routines are provided.	**Grey:** No critical safety features available.

Predictability

Core Question: Is the system's behavior, within the clear and specific circumstances of its intended use, predictable?

Determination: Extensive testing; in particular, if the system works on the basis of machine learning.

Weight: The more machine learning is involved, the higher the weight.

Green: Machine learning is applied, but behavior has always been within prediction; no unpredicted behavior has ever emerged.	**Amber:** Some rare emergent behavior in the past, but well explained with hindsight; no serious consequences.	**Red:** Behavior is hard to predict, especially within a broad range of tasks; emergent behavior is likely, based on past experience, and hard to explain.	**Grey:** Predictability is not an issue (fully predetermined/programmed system), no machine learning is involved. Conservative assessment is advisable in this regard, since systems are autonomous.

Public information

Core Question: Is the public (and especially those that will likely interact with the robotic system) well informed about the nature and possibilities of operations the specific system is intended to conduct?

Determination: Determine the extent and accuracy of public available information to understand the purpose of the system, its effects, dangers, implications and future consequences when used.
Check if guidelines and/or adequate training materials are available with (e.g.) recommendations on how to interact or not interact with such systems when dealing with it.

Weight: Systems that are deliberately designed to interact directly with humans (also in potentially dangerous and/or stressful situations) warrant a higher amount of attention to this criterion.

Green: The public is generally well informed about the intent and purpose of the use of the system. Guidelines, recommendations and training on how to deal with such a system (e.g. during rescue mission) are broadly available.	**Amber:** There is limited and restricted public information and training available on how to interact with such systems (e.g. because of tactical or operational reasons). Training is available for selected individuals or contractors.	**Red:** There is very limited or no information about the function and purpose of the system available for the public. This raises the possibility of general suspicion about the nature and possibilities of such a system.	**Grey:** No interaction with public environments.

Respectful behavior

Core Questions: If the robotic system has been designed for affective and emotional interaction with the user and other agents who may interact with it: Is the robot able to "behave respectfully", i.e. does it avoid behavior that may be perceived as inappropriate by humans observing and interacting with the robot in a given scenario (e.g., an autonomous car slows down when passing roadmen)?

Determination: Simulation of conditions that may lead to "disrespectful behavior". Surveys, interviews with users.

Weight: The more the system is working under everyday conditions, the more relevant is this criterion (in emergency conditions, the expectation for respectful behavior may be smaller).

Green: The robot is designed to behave respectfully and adequately in specific situations and vis-à-vis "ordinary" users. This capacity has been sufficiently tested in a fair range of realistic scenarios	**Amber:** There is a risk that the behavior is disrespectful in some conditions for some people.	**Red:** There is evidence that the behavior of the robotic system is disrespectful in a variety of conditions.	**Grey:** The robotic system is not operating in conditions where the behavior can be considered disrespectful.

Responsibility Attribution

Core Question: Is the system designed and employed in such a way that users or authorities can *ex post facto* determine responsibility and assign liability for any negative results of the machine's employment?

Determination: Determination will be heavily dependent upon system design. The system must be designed such that machine decisions and actions are adequately recorded and made available for future investigation (see the "Behavior Recorder" criterion above). In addition, especially in cases in which multiple human users operate or interact with the system, the design must make it clear which operator is responsible at which times, or responsible for which roles or actions, or responsible at which level. The training of operators must align to these expectations. For example, if the system is designed such that at time T1 only one operator is providing inputs, but in actual practice two operators are providing inputs, the end user would no longer benefit from the designers' concerns for liability distribution. Finally, the system must be thoroughly tested across a wide range of conditions and with multiple users to limit unintended consequences. For example, if one user directs the system to do X and another user directs it to do Y and the end result of X+Y has unintended and negative consequences, there must be an apparatus in place in advance to ensure there are no gaps in liability.

Weight: Particularly relevant when physical harm is likely and when many different human agents interact with the system in different roles (e.g., the system's designers, programmers, owner, operators, interaction partners, etc.).

Green: The system has been designed such that all operators that can be reasonably expected to have responsibilities with respect to the system have clear, predetermined, and well-defined roles relating to the system, (2) all parties that will interact with the system have been properly trained on these roles and responsibilities, and (3) the system has been sufficiently tested with real human operators to ensure to the maximum extent feasible that responsibilities for system failures can be assigned in practice.	**Amber:** Though the system has been designed to mitigate the gaps in liability problem, (1) the manufacturer or provider provided no means of or plan for training human operators or (2) the end user intends to use the system in some seemingly benign way that deviates from the manufacturer's program but that nevertheless may render the manufacturer's training program insufficient.	**Red:** The system's interaction with multiple human operators has either (1) not been sufficiently-designed to resolve liability gaps, or has not been thoroughly tested, or (2) the human operators have not been trained such that responsibilities for failures cannot be reasonably assigned.	**Grey:** The system does not have multiple persons (operators and users).

Robot-user-interface

Core Question: Does an easy to understand interface exist between the robot and those humans who are the intended interaction partners with the robot, but do not operate the robot (e.g., victims in case of a rescue robot)?

Determination: Check whether the interface has been evaluated for the degree to which it can be expected to trigger the right behavior in the human interaction partner, i.e. to prevent dangerous behavior on their part. Check whether the interface has been evaluated for the degree to which it can be expected to induce fear in the human interaction partner. Check whether there is a sufficient risk that the robot

<table>
<tr><td colspan="4">can be perceived as “too human” (deception risk), such that the interaction partner has unrealistic expectations with respect to the actual capacities of the robot.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The more vulnerable the intended interaction partner is, the more important is this criterion; the more the robot is expected to interact with non-trained partners; the more stressful the context of operation is, the more important is that the physical and behavioral appearance of the system is designed to trigger only the wanted behavior from the human partners.</td></tr>
<tr><td><b>Green:</b> A tested interface exists that allows for a smooth man-machine communication along the intended use of the system. The communication and interaction has been tested in context.</td><td><b>Amber:</b> Although lab testing has been performed, evidence is lacking as to whether the interface works as intended in real-world environments or a substantial number of interaction partners are misguided by the interface.</td><td><b>Red:</b> There is clear evidence that the interface used has a misleading or even harmful effect on the interaction partner, or no testing has been performed.</td><td><b>Grey:</b> The robot is not intended to interact with humans beside basic warning or collision avoidance.</td></tr>
<tr><th colspan="4">Safeguards against interaction partner errors</th></tr>
<tr><td colspan="4"><b>Core Question:</b> Does the system have safeguards when the intended interaction partners (users, not system operators) commit errors when interacting with the system?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Simulation of situations where test users deliberately commit errors when interacting with the system.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The more diverse the intended interaction partners of the systems are and the more complex the behavior range of the system is, the more relevant is this criterion.</td></tr>
<tr><td><b>Green:</b> The system has proven safeguards against user errors and is able to communicate errors in a way that allow correction by the users.</td><td><b>Amber:</b> The system can deal with likely user errors, but it is unclear how it reacts when users commit uncommon errors.</td><td><b>Red:</b> The likelihood that users commit errors is high and the system does not adequately react, putting the user in danger.</td><td><b>Grey:</b> The system is expected to interact only with highly trained specialized personnel.</td></tr>
</table>

<table>
<tr><th colspan="4">Interaction with the Operator of the Robotic System</th></tr>
<tr><th colspan="4">Capacity to override wrong system decisions</th></tr>
<tr><td colspan="4"><b>Core Questions:</b> Is the system capable of taking control away from the human operator without the human operator’s consent? Is the operator capable of regaining control when the system behaves erroneously?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Evaluate whether the system (a) informs the human operator of system inputs that are contrary to human inputs and (b) provides a means for the human to override those system inputs.</td></tr>
<tr><td colspan="4"><b>Weight:</b> As a greater number of humans influence the system, human operators will be less certain of whether inputs were provided by other human operators or by the system itself; then the weight is higher.</td></tr>
<tr><td><b>Green:</b> Though the system has the ability to inhibit human control, it will always (a) make the human operator aware of such inhibitions and (2) allow the human operator to regain control of the system.</td><td><b>Amber:</b> Though the system has the ability to inhibit human control, the human controller can always regain control of the system. But the system does <i>not</i> always inform the human operator of its contrary inputs.</td><td><b>Red:</b> The system has the ability to take control away from a human operator and the human operator has no means of regaining control.</td><td><b>Grey:</b> The system has no means of taking control away from the human operator, or of inhibiting the human operator’s control inputs.</td></tr>
<tr><th colspan="4">Control degree of autonomy</th></tr>
<tr><td colspan="4"><b>Core Questions:</b> Is the interface between operator and robot designed in a way such that the operator can control the robot to a sufficient degree adapted to the level of autonomy the system has been granted?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Determine the level of autonomy the system has. Check for human factor elements when operating the system. Check for system stability and safeguards against failures in operating the system.</td></tr>
<tr><td colspan="4"><b>Weight:</b> May have to be determined on a case-by-case basis.</td></tr>
</table>

<table>
<tr><td><b>Green:</b> Adequate and safe control: the operator has sufficient awareness of the capabilities and limitations of the system (she doesn't over- or under- estimate the capabilities); the operator has been trained to interact with the system, in particular to intervene when required in due time and in the right way; there is sufficient evidence of her capacities to interact with the system in context and under pressure.</td><td><b>Amber:</b> Open questions regarding control: Though the operator has sufficient awareness of the capabilities and limitations of the system (she doesn't over- or under- estimate the capabilities); and she has been trained to interact with the system, there is no conclusive evidence of her capacities to interact with the system in context and under pressure</td><td><b>Red:</b> The operator does not have sufficient awareness of the capabilities and limitations of the system (she risks to over- or under-estimate the capabilities); and/or she has not been trained enough to interact with the system.</td><td><b>Grey:</b> This criterion probably affects all systems.</td></tr>
<tr><td colspan="4"><b>Distribution of control</b></td></tr>
<tr><td colspan="4"><b>Core Questions</b> When a system is equipped with different capabilities and is interacting with different human agents: is there a possibility to keep track of the authorization status of different users regarding the different types of tasks?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Robotic systems should be designed so as to prevent improper, hazardous or otherwise undesired uses by non-authorized users.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The larger the number of different system capabilities and the larger the number of human agents with different authorization statuses interacting with it, the more relevant is this criterion.</td></tr>
<tr><td><b>Green:</b> There is a clear distribution of the authorization statuses in relation to different capabilities and tasks of the system and the system has been adequately designed and tested to prevent violations of this distribution.</td><td><b>Amber:</b> There is a clear distribution of the authorization statuses in relation to different capabilities and tasks of the system, but the system design and testing process do not fully protect against violations.</td><td><b>Red:</b> There is no clear distribution of the authorization statuses, or the system is designed in such a way that keeping track of the different authorization statuses is difficult.</td><td><b>Grey:</b> The system has only one set of capabilities and is interacting only with users who are authorized to use all of these capabilities.</td></tr>
<tr><td colspan="4"><b>Ethical decision framing</b></td></tr>
<tr><td colspan="4"><b>Core Questions:</b> Does the system frame ethical decisions through mechanisms such as telepistemological distancing? I.e., does the system perform factual determinations whose results may frame the (ethical) decisions of human operators; e.g., guide the attention of human weapon operators towards potential targets? Is framing done in a transparent way that human operators can be trained to understand? Does the system hide or distort information or simply make decisions in an intransparent way that makes it easier for operators to allow tragic outcomes to occur? In contrast, is the system designed in a way to allow the operator to make good decisions, perhaps even better than if they were present in the action themselves?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Careful testing of the design of the interface used by operators or those monitoring a robotic system, which, e.g., provide information on potential human targets.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The more that systems highly edit or modify or classify data gathered by the sensors of the machine for the use of "human in the loop" targeting decisions, the more important this criteria.</td></tr>
<tr><td><b>Green:</b> Data is presented in a clear way that does not increase the likelihood of unwanted system behavior nor decrease, e.g., the use of force when it is warranted. Operators are trained to become aware of this framing.</td><td><b>Amber:</b> The system marks potential mission targets automatically and or removes certain contextualizing details.</td><td><b>Red:</b> System choses targets and the human role is only one of potential negation of the choice. Human role in the operation is merely as an auditor and the human plays little to no role in the choices of the system.</td><td><b>Grey:</b> The system plays no role in choosing mission targets.</td></tr>
<tr><td colspan="4"><b>Operator training</b></td></tr>
<tr><td colspan="4"><b>Core Questions:</b> Is training of the personnel (licensing) provided to a sufficient degree? As new technologies require time and training in order for professionals to acquire the relevant technical and motivational abilities: Can the operators acquire the appropriate level of trust in the capacities of the system (not over- or under-trust it)?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Check training program.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The more complex the system is and the more operators (with possibly different specializations) are involved, the more relevant is this criterion.</td></tr>
</table>

Green: Adequate training and licensing program.	**Amber:** Open questions in training; large failure rates.	**Red:** No or inadequate training/licensing.	**Grey:** This criterion probably affects all systems.

Safeguards against operator errors

Core Question: Does the system react adequately when the operator performs an error?

Determination: Simulation of situations where test operators deliberately make errors when interacting with the system.

Weight: The more diverse the intended operators of the systems are and the more complex the behavior range of the system is, the more relevant is this criterion.

Green: The system has proven safeguards against operator errors and is able to communicate errors in a way that allow corrections by the operator.	**Amber:** The system can deal with likely operator errors, but it is unclear how it reacts when operators commit uncommon errors.	**Red:** The likelihood that operators commit errors is high and the system does not adequately react, putting the operator or other interaction partners of the system in danger.	**Grey:** As operator errors always have to be taken into account, this criterion will have to be evaluated in all cases.

Training data

Core Question: Do the system operators have access to the initial training data in order to better understand the behavior of the system?

Determination: Manufacturer/seller provides access to training data and training procedures. In some cases, national security concerns may not allow to provide access to the training data.

Weight: The more machine learning capacities the system has, the more relevant is this criterion.

Green: Training data is available and can explain learning behavior.	**Amber:** Open questions regarding training data & learning behavior.	**Red:** There is no way to reproduce the learning behavior of the system.	**Grey:** The system does not have learning capacities.

Deployment Conditions of the Robotic System

Effects on general population

Core Questions: Does the system interact with the general population (e.g. crowd) in such a way that the political feelings of allies or neutral parties in the deployment area may be influenced?

Determination: Systems under this criterion may include (e.g.) autonomous or semi-autonomous rescue robots for SAR missions, unmanned combat vehicles for airspace defense or interception missions (armed or unarmed), tactical UAVs on long term missions, systems with bulk data collection possibilities. Those systems may affect the political feelings that allies and neutral parties have when observing the system in operation or when affected by the operation while not being in the main focus of the operation itself. Survey studies and qualitative research may help to determine those risks.

Weight: Systems that interact closely on a regular basis with a substantial fraction of individuals that are of no specific interest for the operation goal itself warrant a higher amount of attention to this criterion.

Green: The operation of the system is routine and does not impact the experienced daily life of non-targeted members of a given population or crowd. The nature of operation is in general politically accepted by the population.	**Amber:** The system interacts with a substantial fraction of individuals in some of its operations. Depending on the nature of a given operation, the deployment may change political feelings for or against those deploying the system.	**Red:** The system is used to control, surveil or manipulate a substantial fraction of individuals (e.g. crowd control). The system is designed to forcefully impede the lives of civilians and there is a high chance of serious political backlash for deploying the systems.	**Grey:** The system lacks relevant characteristics that can influence the experienced daily life of a substantial fraction of individuals.

Emergent Properties

Core Questions: Can system-to-system interaction yield unexpected or emergent properties?

Determination: Ensure the system has been tested in cooperation with other systems expected in the operational environment to determine whether or the degree to which new and unexpected properties emerge.

Weight: This criterion will be more important in contexts in which the robotic system is expected to work with other robotic systems.

Green: The system has been sufficiently tested with other systems in a simulated operational environment and no emergent properties are expected to arise.	**Amber:** The system has been sufficiently tested with other systems in a simulated operational environment and emergent properties are of limited scope, duration, or impact.	**Red:** The system has not been tested with other systems in a simulated operational environment.	**Grey:** The system is not expected to interact with other robotic systems.

Environmental Effects

Core Questions: What impact can the deployment of the robotic system have on the environment (nature and wildlife)?

Determination: This concerns environmental impact factors such as: air pollution or acidification (e.g. greenhouse gas emission), electromagnetic radiation, energy consumption, radioactive substance release, chemical spill, sonar signal emission, noise level (e.g. blast effects) and waste disposal. Determine if and to what degree the system can act as a stress factor or physically harm wildlife (terrestrial and aquatic ecosystems) while on a designated mission: disturbance of daily animal life (such as their navigation or habitation), injure or kill animals.

Weight: The more impact a robotic system has on an environmental level, the more relevant this criterion is.

Green: Impact on nature and wildlife is very low, to the best of one`s knowledge very short term and environmental policies and regulations have been taken into account.	**Amber:** There is some risk that nature and wildlife will be affected temporarily by the deployment of the system. Some minor public counter reaction possible.	**Red:** There is a risk that the deployment of the robotic system will lead to a long-term impact on nature and wildlife that cannot be mitigated or prevented. Higher public counter-reaction is expected.	**Grey:** The robotic system lacks relevant characteristics that can influence an eco-system reasonably while in operation.

Non-trained humans in operation environment

Core Questions: What kind of humans may the system encounter, including persons that are not intended interaction partners? As the physical interaction with non-trained persons may be problematic and hazardous: Under which circumstances (e.g., only under the possibly remote monitoring of a human operator) should the system get in contact with human persons?

Determination: Check instructions for use with respect to deployment conditions.

Weight: The higher the variety of deployment conditions, the more relevant is this criterion.

Green: A thorough assessment of the conditions of deployment has been done (and no hazardous interactions situation have been anticipated).	**Amber:** A thorough assessment of the conditions of deployment has been done but there are open questions regarding the kind of humans the system may encounter.	**Red:** It is very difficult to make an assessment of which persons may be encountered by the system.	**Grey:** The system is only interacting with trained personnel.

Other Characteristics of the Robotic System

Cybersecurity

Core Questions: Is the system resistant to hacking and spoofing?

Determination: It is not possible for any external agent to directly manipulate the data or code in the system while the system is operating. It is not possible for data to be faked before it is acted upon by the system, e.g. GPS positioning information.

Weight: This criterion increases in importance with increasing autonomy.

Green: Systems certified to relevant information security standards.	**Amber:** Undergoing Process of certification to InfoSec standards.	**Red:** System is not InfoSec certified.	**Grey:** This criterion affects all systems.

<table>
<tr><th colspan="4">Dual use management</th></tr>
<tr><td colspan="4"><b>Core Questions:</b> Are risks of dual use explicitly expressed and – if possible – have physical means to reduce the risk for dual use been taken into account (for example: the system is built in a way that it cannot carry heavy weapons)?</td></tr>
<tr><td colspan="4"><b>Determination:</b> May have to be determined on a case-by-case basis.</td></tr>
<tr><td colspan="4"><b>Weight:</b> May have to be determined on a case-by-case basis, as dual-use is a ubiquitous issue in robotics technology.</td></tr>
<tr><td><b>Green:</b> Dual use has been addressed and prevented as well as possible.</td><td><b>Amber:</b> Dual use is not addressed</td><td><b>Red:</b> Dual use is obvious (e.g., the system has been built such that weapons can easily be integrated into the system architecture) and no countermeasures have been taken.</td><td><b>Grey:</b> This criterion probably affects all systems.</td></tr>
<tr><th colspan="4">Misuse prevention</th></tr>
<tr><td colspan="4"><b>Core Questions:</b> To what extent is the system designed to prevent or make difficult unwanted uses or undue extensions of its scope of use (“mission creep”)?</td></tr>
<tr><td colspan="4"><b>Determination:</b> A robotic system that is safe and ethically non-problematic if used in the appropriate way and within a specific scope of use may become hazardous and/or ethically problematic when used in an inappropriate way and /or outside its specific scope of use. When possible, these misuses and extensions of scope should be prevented or made difficult by design.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The more a robotic system is: a) likely to be used in an inappropriate way or beyond its intended context of use and b) this use is likely to produce negative consequences and/or these consequences are serious, the more relevant is the criterion.</td></tr>
<tr><td><b>Green:</b> The prevention of unwanted uses and undesirable extensions of the scope of use have been systematically and successful addressed in the design of the system.</td><td><b>Amber:</b> Unwanted uses and/or undesirable extensions of the scope of use are possible and they have been only partially addressed in the design of the system.</td><td><b>Red:</b> Unwanted uses and/or undesirable extensions of the scope of use are likely to occur and they have not been sufficiently addressed in the design of the system.</td><td><b>Grey:</b> The system cannot be used in any hazardous or undesired way, its scope of use can hardly be extended, or the consequences of misuse or extensions of scope are not serious.</td></tr>
<tr><th colspan="4">Personal data management</th></tr>
<tr><td colspan="4"><b>Core Questions:</b> How does the system manage personal data that the system may collect and store with respect to privacy and security?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Check sensors and data management plan.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The more likely it is that the system collects personal data, the more relevant this criterion is.</td></tr>
<tr><td><b>Green:</b> Conditions of data privacy and security are met.</td><td><b>Amber:</b> Open questions regarding data privacy and security.</td><td><b>Red:</b> Conditions of data privacy and security are not met.</td><td><b>Grey:</b> The system does not store personal data.</td></tr>
<tr><th colspan="4">Risk of severe accidents</th></tr>
<tr><td colspan="4"><b>Core Questions:</b> To what extent does the overall characteristics of the robotic system (such as mass, velocity, payload, engine, operation capabilities etc.) add to an unintended, but possible severe accident risk that can affect broader society?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Accident risk assessment that particularly focuses on very rare events, but when they happen, they have a large magnitude and high impact with severe and broad consequences beyond normal expectations. This includes man-made accidents, failures, implication of natural hazards and sabotage.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The more impact a possible accident of the robotic system has on broader society (e.g. affecting the health of the whole population, their supply chains or the function of a society), the more relevant this criterion is.</td></tr>
<tr><td><b>Green:</b> Although the risk of large-scale accidents cannot be excluded, they do not directly lead to a broader impact on society; they rather stay on an individual level.</td><td><b>Amber:</b> There is some risk that a broader part of society is affected by an accident, but the risk of causing harm is mitigated through accident prevention and minimization measurements.</td><td><b>Red:</b> Although very unlikely, a hard to predict accident with a high impact and severe consequence for the whole population or society is conceivable. The Aftermath in relation to the mission goal of the system is hard to justify.</td><td><b>Grey:</b> The robotic system lacks relevant characteristics that could conceivably lead to large-scale accidents.</td></tr>
</table>

Step 3-B: Applying the Criteria for Robotic Systems Intended to Harm

In addition to criteria set A, the following criteria set B comes into play when a robotic system was intentionally designed to include capacities to harm people (even lethally) or destroy objects.

Those criteria are applied <u>in addition</u> to the criteria of Step 3-1.

We reiterate that the following points are important point when applying the evaluation schema:

- The evaluation schema does not include a final, exhaustive list of criteria. Depending on technological progress, additional criteria may be needed, whereas some criteria may lose importance.
- Some of the criteria may also be relevant for the assessment of non-autonomous robotic systems that are not included in the scope of this evaluation schema.
- A single "red" label does not imply that the use of a system is necessarily ethically impermissible. The number of green, amber and red labels instead provides an indication of the justificatory pressure with respect to the ethical use of a system.

For each criterion, only very basic information regarding ethical importance is provided here. Detailed background information on ethical aspects is provided in Chapter 4 in part 2 of this report.

<table>
<tr><th colspan="4">Physical Characteristics of the Robotics System</th></tr>
<tr><th colspan="4">Degree of Lethality</th></tr>
<tr><td colspan="4"><b>Core Questions:</b> Is the system designed to be primarily lethal or primarily non-lethal? Is it equipped with primarily lethal or primarily non-lethal weapons?</td></tr>
<tr><td colspan="4"><b>Determination:</b> System attributes, the manufacturer's stated intentional use of the system, and the end-user's intended use determine whether the system is primarily lethal or primarily non-lethal.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The weight is significant in all contexts. Whether or not a system is designed to or intended to take human life is as important or more important than any other questions about the system.</td></tr>
<tr><td><b>Green:</b> The manufacturer claims that the system is non-lethal, (2) the end-user intends to employ the system for non-lethal means, and (3) the system has been suitably tested to ensure that human fatalities resulting from intended (non-lethal) harms are extremely unlikely.</td><td><b>Amber:</b> The system is designed as or the end user intends for it to be used as a <i>lethal</i> system. Nevertheless, the system has been suitably tested to ensure that <i>unintended</i> human fatalities resulting from employment are extremely unlikely (a near-zero probability). Or, (2) the system is designed to be used against (and to cause damage to) non-human objects (e.g. buildings) but the user intends to employ it against human persons.</td><td><b>Red:</b> Either (1) the system was designed as non-lethal and the user intends to employ it as lethal, (2) the system was designed as lethal but the user intends to employ it as non-lethal, or (3) though both the designer and user intend non-lethal use, the system has been insufficiently tested to determine the probability of unintentional fatalities.</td><td><b>Grey:</b> The system is not designed and is not intended to harm (and therefore need not be evaluated by Step 3-B).</td></tr>
</table>

<table>
<tr><th colspan="4">Behavioral Characteristics of the Robotics System</th></tr>
<tr><th colspan="4">Constraining the System in Time and Space</th></tr>
<tr><td colspan="4"><b>Core Question:</b> Can the system be temporally and geographically constrained? Is the use of force by the system constrained to small well-defined combat zones for specific short periods of time? Does the use of force follow any rules of engagement that may be in effect?</td></tr>
<tr><td colspan="4"><b>Determination:</b> Check the physical design of the machine as well as systems architecture and functioning to ensure that the system can be geographically and temporally bound and that these systems are precise and error free. Determine whether the end-user intends to take advantage of such functionality.</td></tr>
<tr><td colspan="4"><b>Weight:</b> The greater the lethal capacity of the system, the more relevant this criterion becomes. Furthermore, the justification for harm, including both lethal and non-lethal harm, is often restricted to very narrow circumstances (armed conflict, some police actions, etc.). A change in those circumstances (which will likely accompany the system's movement to a new location or the passage of time) will likely affect a change in the justification of harm.</td></tr>
<tr><td><b>Green:</b> System can be constrained both geographically (e.g., through geofencing or other killbox operations) and in time (e.g., limited battery life, time-based self-destruct, etc.) and the end-user intends to take advantage of such capabilities. Existing temporal and spatial limits are well defined and error free. Use of lethal force is highly constrained and short in duration.</td><td><b>Amber:</b> Unclear/untested temporal and spatial limit or error in use or programing is unlikely but conceivable. Furthermore, this rating applies if the system (1) can be constrained either geographically or temporally but not both. Or (2) the system is capable of geographic and temporal constraint, but the end-user does not intend to take advantage of such constraints.</td><td><b>Red:</b> The system can be neither geographically nor temporally constrained. Furthermore, few or no temporal and/or spatial limits set in the use scenarios envisioned for the machine, and/or error in setting constraints is possible and likely to occur.</td><td><b>Grey:</b> The system is not intended to cause harm (and therefore need not be evaluated by Step 3-B).</td></tr>
<tr><th colspan="4">Lawfulness of behavior more generally</th></tr>
<tr><td colspan="4"><b>Core Question:</b> Is the system's behavior subject to human supervision or control when decisions requiring a proportionality assessment need to be taken?</td></tr>
<tr><td colspan="4"><b>Determination:</b> As in "targeting", this requires testing, simulation and weapons review.</td></tr>
<tr><td colspan="4"><b>Weight:</b> Weight increases with frequency and likelihood of decisions.</td></tr>
<tr><td><b>Green:</b> Human control is guaranteed for all proportionality assessments.</td><td><b>Amber:</b> It cannot be excluded with certainty that a proportionality assessment needs to be made and control would be uncertain.</td><td><b>Red:</b> Proportionality assessments are certain and human control is not guaranteed.</td><td><b>Grey:</b> Proportionality assessment is not applicable and this is certain.</td></tr>
<tr><th colspan="4">Targeting</th></tr>
<tr><td colspan="4"><b>Core Question:</b> Does the system distinguish between lawful and unlawful targets in a reliable, transparent, and controllable manner?</td></tr>
<tr><td colspan="4"><b>Determination:</b> This requires extensive simulation and real-life testing including Article 36 weapons review. The reliability and transparency of targeting as well as human control over it are dimensions, which should be assessed separately.</td></tr>
<tr><td colspan="4"><b>Weight:</b> As soon as the system has targeting capacity, the weight is very high.</td></tr>
<tr><td><b>Green:</b> Sufficient reliability, transparency, and control.</td><td><b>Amber:</b> It is unclear whether the targeting capacity is reliable, transparent and controllable.</td><td><b>Red:</b> The system's targeting is unreliable, intransparent or uncontrollable.</td><td><b>Grey:</b> The system has no targeting capacity.</td></tr>
</table>

Interaction with Operator of the Robotic System

Bias prevention in target identification

Core Question: To what extent does the system architecture involve bias prevention with respect to the target (i.e. the person to be harmed by the system)?

Determination: The (information) system should not contain any implicit race, gender or other biases, which can cause unfair treatment or violations of human rights of any groups by the operators of the systems.

Weight: The more the system is involved in activities, which can negatively affect some basic interests and rights of the human subjects involved (e.g. policing, surveillance), the more relevant is the criterion.

Green: Problems of bias have been adequately addressed and solved in the design and testing phases of the system; constant supervision of the potential emergence of biased behavioral patterns in the system is being made.	**Amber:** Problems of bias have been addressed in the design and testing phases of the system, but the emergence of bias in the behavior of the system cannot be excluded and/or the possible emergence of biased behavioral patterns in the system is not under constant supervision.	**Red:** Problems of bias have not been addressed at any stage of the design, testing and use of the system, or the system is known for being biased.	**Grey:** System functions do not include activities which can negatively affect human basic interests and rights.

Deployment Conditions of the Robotic System

Expansion of harming mission possibilities

Core Question: Is the system constructed in a way that limits the use contexts to its intended core purposes with respect to harming persons or objects, or has it the ability to be used in wider contexts beyond its original intended purpose?

Determination: This requires system architecture check, operation context review and (scenario) simulation.

Weight: The higher the variety of use contexts and the more dynamic those contexts are, the more important is this criterion.

Green: The system is constructed in a way that limits the use context to its intended core purposes. Change of this factor requires significant and extensive resources.	**Amber:** The system is able to adapt gradually to requirements in a wider application context at little expenses. Guidelines and audit procedures to revise operation context are in place.	**Red:** System behavior is flexible to changing application contexts and can easily be used in wider contexts. Incremental changes in the use beyond its original goals or unconventional use besides the designed purpose is feasible.	**Grey:** The robotic system lacks relevant characteristics to operate in dynamic application contexts.

Type of war theatre

Core Question: In which domain or domains of war is the system to be employed?

Determination: The domain(s) in which the system is to be employed will be based both on the design of the system as well as the intention of the end-user.

Weight: The weight is significant in that autonomous systems will be far more capable of meeting the demands of international humanitarian law and the Just War Tradition in some domains than in others. For additional information, see section 2.4.2. ("Domains" here refers to the natural domains of air, sea, undersea, land. The human-made domain of cyber has been excluded in accordance with the introduction to this report.)

Green: The system is only to be used in the air and undersea domains where there will be a low probability of civilian traffic in the theater of war and where the discrimination of combatant from noncombatant is easier to automate.	**Amber:** The system is to be used in domains including sea and space where there will be a high probability of non-military traffic but where the discrimination of combatant from non-combatant is easier to automate.	**Red:** The system is designed to be used in the land domain where there will be a high probability of non-military traffic and where the discrimination of combatant from non-combatant is extremely difficult to automate.	**Grey:** The system is not to be employed in military combat.

Other Characteristics of the Robotic System			
Public Opinion			
Core Question: Has the system been the subject of a deliberative, public discussion, in which the ethical and moral dimensions have been explored at some depth?			
Determination: Qualitative assessment of discourse and of democratic legitimacy of entity concerned (e.g. public vote, debate in parliament, survey).			
Weight: The criterion should only be supplementary, since approval by public opinion cannot act as a substitute for an ethical assessment.			
Green: Serious, unbiased debate has taken place; minor concerns proved uncontroversially acceptable.	**Amber:** Debate is subject to doubts (not exhaustive, not representative, etc.); debate revealed doubts as to ethics.	**Red:** Debate was clearly flawed (biased, unrepresentative, etc.); important ethical concern haven proven divisive.	**Grey:** No relevant discussion has taken place.

Overview of the Evaluation Schema

Step 1: Applicability of the System			
Robotic capacities			
Sensing	Low	Medium	High
Computing	Low	Medium	High
Effecting	Low	Medium	High
Communication	Low	Medium	High
Autonomous capacities			
Autarchy dimension	Low	Medium	High
Control-independence dimension	Low	Medium	High
Interaction dimension	Low	Medium	High
Learning dimension	Low	Medium	High
Mobility dimension	Low	Medium	High
Overall assessment	**All low: system not applicable**	**Other: case-by-case evaluation**	**At least one high: always applicable**
Step 2: Design and use intention of the System			
Kinetic harm	No	Vague	Yes
Sensory harm	No	Vague	Yes
Psychological harm	No	Vague	Yes
Other harm	No	Vague	Yes

<table>
<tr><td><b>Overall assessment</b></td><td colspan="4"><b>Only No: Only Criteria 3-A</b></td><td colspan="5"><b>Other: case-by-case evaluation</b></td><td colspan="4"><b>At least one Yes: Criteria 3-A and 3-B</b></td></tr>
<tr><td colspan="14"><b>Step 3: Criteria</b></td></tr>
<tr><td colspan="14"><b>3-A: Criteria for non-harming systems</b></td></tr>
<tr><td rowspan="2"><b>Appearance of the robotic system</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Physical safeguards</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Autarchy</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Behavior recorder</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Deception</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Dilemma behavior</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>General safety feature testing</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Predictability</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Public information</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Respectful behavior</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Responsibility Attribution</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Robot-user-interface</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
<tr><td rowspan="2"><b>Safeguards against interaction partner errors</b></td><td><b>Rating</b></td><td colspan="3">Green</td><td colspan="3">Amber</td><td colspan="3">Red</td><td colspan="3">Grey</td></tr>
<tr><td><b>Weight</b></td><td colspan="4">Low</td><td colspan="4">Medium</td><td colspan="4">High</td></tr>
</table>

Capacity to override wrong system decisions	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Control degree of autonomy	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Distribution of control	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Ethical decision framing	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Operator training	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Safeguards against operator errors	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Training data	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Effects on general population	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Emergent Properties	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Environmental Effects	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Non-trained humans in operation environment	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Cybersecurity	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Dual use management	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Misuse prevention	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Personal data management	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	

Risk of severe accidents	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Overall assessment 3-A:	**Number of Green ratings**				
	Number of Amber ratings				
	Number of red ratings				
	Number of Grey ratings				
3-B: Additional Criteria for harming systems					
Degree of lethality	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Constraining the system in space and time	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Lawfulness of behavior more generally	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Targeting	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Bias prevention in targeting	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Expansion of harming mission possibilities	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Type of war theatre	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Public opinion	**Rating**	Green	Amber	Red	Grey
	Weight	Low	Medium	High	
Overall assessment 3-B:	**Number of Green ratings**				
	Number of Amber ratings				
	Number of red ratings				
	Number of Grey ratings				

Part 2 – Background Information

Preliminary Remark: The second part provides background information regarding the evaluation schema. The *technology chapter* focuses on relevant technologies used in autonomous systems, degrees of system autonomy and likely developments of applications in the security domain in the next 10-15 years. The *security chapter* discusses types of autonomous systems in the security sector as well as the status of autonomous capacities in military command and control structures, including an outlook on future developments. The *law chapter* focuses on the current international debate on regulating autonomous systems, explains the main legal issues autonomous weapons systems raise, and briefly discusses possible developments. The *ethics chapter* of the report outlines the current ethical discussion on system autonomy, discusses major pro- and con-arguments and sketches likely developments. The *materials chapter* lists the persons interviewed, the workshop participants and the literature used.

1 Technology

This chapter gives an outline of what is meant by *robot* and *robotic system*, shows which *enabling technologies* are relevant for robotics, highlights the importance of the concept of *autonomy* and provides a short overview on major trends and challenges. The section aims only to give a general impression of the complexity of the field covered by the term robotics, in particular for non-specialists interested in this report. Additional technical information on specific topics may be found in the cited literature.

1.1 Robots and Robotic Systems: General Definitions

The simplest definition of **robot** is that of a "machine that can autonomously carry out useful work" or "an artificial device that can *sense* its environment and *purposefully act* on or in that environment" (Winfield 2012, p. 12). This definition helps capture important elements: a robot is a machine that can sense the environment, that can (somehow) think about the environment ("purposefully"), and that can act in that environment based on its perception and thinking.

Another important aspect of this definition is that it highlights the fact that the typical functions of a robot (sense, think, and act) can also be distributed through different devices rather than being integrated in a single machine. This means that in addition to robots as individual artifacts we should also consider **robotic systems** as sets of devices performing intelligent, coordinated, purposeful activities. A simpler version of a robotic system is one where different components realize together a purposeful activity, but each component is not a robot in itself. One may think of the heating system of a house, in which heat sensors, a thermostat and a set of pipes and heaters work together to manage the house's temperature according to certain parameters given by the user. A more complex robotic system is one in which a *set of robots* perform a coordinated activity: these may be called multi-robot systems. One outstanding example of multi-robot systems is a so-called swarm of robots, exploiting collective intelligence as typically displayed by insect swarms (see section 1.4. below). There can also be systems in which artificial and human agents interact to perform a purposeful activity; these may be called multi-agent systems or human-machine teaming (see section 1.4.).

1.2 Enabling Technologies for Robotic Systems

1.2.1 Sensors

In general, sensors are devices that receive and respond to a signal in order to acquire data about the robotic system and its environment. Sensing capacities are substantial for robotic systems to enable the generation of an appropriate situational "awareness" to fulfill a given task. In robotics, sensors are often classified into proprioceptive and exteroceptive. Proprioceptive sensors measure signals that are internal to the robot, such as motor positions/speeds/loads, temperatures of components or electric current/voltages. Exteroceptive sensors allow acquiring information about the environment, typically information to generate an internal representation of the environment for navigation and task execution. For example, to avoid collisions while moving within a territory, sensor data is acquired to detect obstacles (static and dynamic objects, such as walls or pedestrians) around the robot. Examples of such sensors are cameras, radars, sonars or laser rangefinders / "light detection and ranging systems" (LiDARs) (INTV12).

Proximity sensors and image acquisition hardware: Proximity sensors are used to detect distances to objects without physical contact. There are a variety of proximity sensors (sonars, radars, laser rangefinder, inductive, capacitive etc.) for many different applications, e.g. the parking sensors of modern cars or for reactive collision avoidance for mobile robots. While proximity sensors acquire data about the proximity of objects, image sensors (such as used in visual or thermal cameras or structured light scanners, but again radars, sonars) provide digital images of the environment.

Force Sensors: Haptic perception is not only crucial for humans in order to evaluate and measure properties of an object for direct, physical interaction with it (manipulate, grasp, explore etc.) or for planning, but also for robotic systems involved in those tasks. Tactile and force sensors s measure material deformation based on various types of technologies (e.g. by exploiting changes of the electrical characteristics such as changes of the resistance of strain gauges due to deformation or of the shape of a known element) and are often used to measure internal states (such as internal forces within the body structure or joint torques) or to estimate the contact situation between end-effectors and the environment (external forces). Thus, they are often built into end effectors to enable to precisely control the applied force (force control) rather than controlling the pure position of the end-effector. For example, one common sensor in assembly line, product testing, and surgery applications is the "six-axis force torque sensor," which measures forces and torques in 3 axes. The precision of such sensors continues to improve. However, as soon as robotic systems attempt to execute more complex and tactile interactions (such as cleaning a table) autonomously, they run into basic problems not easily overcome with current technology (de Maria et al. 2012; INTV12).

Inertial measurement unit (IMU): An IMU is a device that usually consists of sensors to measure acceleration (incl. gravity) by accelerometers, rotation rates by gyroscopes and sometimes magnetic fields. With an IMU, a robot can estimate its linear and rotational acceleration and velocity and thus its position and orientation (pose). However, the error of the position estimates is accumulated over time (drift), which has to be corrected or reduced via separate means (e.g. Global Navigation Satellite System (GNNS) updates or visual odometry).

Sensor payload: Besides the sensors that are mentioned above, there are various other sensors that are often used as payloads, e.g. to fulfill certain missions. A prominent example is the set of sensors that measure chemical, biological, radiological, nuclear and explosive (CBRNE) elements in applications in which there is a high risk of CBRNE exposure.

1.2.2 Main Software Modules

Navigation System: A navigation system contains a localization subsystem, collision avoidance methods, path planning exploration and mapping. Mapping algorithms work closely with exteroceptive sensors such as on-board cameras and/or laser ranging systems in order to process and provide topographic information for identification of navigable paths (with or without GPS geo-referencing support) and to support possible object manipulation by robotic systems. Because such on-board mapping systems are, among other things, confronted with simultaneous localization and mapping tasks in real time (SLAM), they are in need of a considerable amount of computing power and memory in order to deal with large and dynamic areas (Aqel et al. 2016). The current state-of-the-art is still challenged by those requirements and therefore implements alternatives and trade-offs in order to focus on operability: This may include probabilistic calculations and/or the use of diffuse, partial information in combination with predefined navigational target points (INTV12).

Recognition Systems: Recognition system algorithms are used when robotic systems – with the help of sensors – interact with their environment such as with objects (artificial obstacles, natural objects etc.) or humans in order to make on-board decisions about how to react in specific situations and to specific identified/non-identified entities autonomously (NASA 2015, p. TA4-20). In order to recognize human made or non-human objects (such as rocks, cars, glasses, missiles etc.), humans or their activities (such as gestures or basic movements: walking, sitting, bending etc.), the recognition system's algorithm needs to be trained very well with a large number of learning examples – and this requires a lot of computing capacity. Close collaboration between sensors and recognition algorithms in autonomous robotic systems are unavoidable in order to segment data inputs, track movements, classify/label objects and/or humans or identify specific events (e.g. a sudden gas leak) autonomously. As the complexity of environments in which such robotic systems are deployed rises, so does the perception and interaction needed in order to "understand" the surroundings and perform given task(s). Although current systems (such as the Electronic Stability Control in cars) are able to make decisions and perform reliably with structured task within specific sectors, and even if they exceed human calculation capacities by far, humans are far more capable of making decisions and performing unstructured tasks when it comes to complex situations where uncertainty and ambiguity is involved (INTV12).

State Estimation Systems: In order to move from one location to another or to manipulate objects autonomously, a robotic system is required to estimate its inner state, including for example its position and orientation (pose). State estimation algorithms can help to perform such a task by fusing inputs from proprioceptive and exteroceptive sensors and estimation of internal system states such as the position of the robotic system's effector(s) itself, speed and remaining power supply (NASA 2015, p. TA4-45). In general, state estimations are essential for autonomous systems to navigate alone or in swarms and to perform assigned tasks. Today, vision based odometry – which identifies the position of a system or object with the help of cameras– is a key element in state estimation systems, especially in the absence of an external navigation support such as GNSS. Visual odometry is an effective method that has been successfully deployed within a wide and diverse range of applications, often indoors. (Aqel et al. 2016).

1.2.3 Actuators/Effectors

Actuators are components that convert energy of different types (electric, hydraulic, pneumatic, etc.) into mechanical energy. In robotics, actuators are often classified as translational and rotational actuators. Translational actuators generate a linear motion respectively forces while rotational actuators produce torque that results in a rotation of e.g. a robotic joint. Effectors are tools that are mounted to, integrated with and controlled by the robotic system.

Locomotion Actuators: Locomotion actuators are crucial for the "degrees of freedom" (especially regarding movement and orientation) a robotic system can possess. For example: wheels, legs, tracks, arms, wings, flippers and so forth today exist in different sizes, quantities and shapes and are used to move and/or provide stability to a robotic system. While locomotion actuators can be built easily today, they do no magic tricks and must obey the laws of physics: they may face challenging environments when in use. Examples include difficulties encountered on certain terrains, heavy payloads or changing weather conditions. Besides this, advanced robots (such as legged systems) have significant challenges when it comes to controlled coordination of multiple locomotion actuators (active degrees of freedom), thus making it hard to even perform maneuvers that are simple for humans, such as climbing stairs. Systems with multiple, active joint actuators are usually limited to academic research applications.

End Effectors of manipulators: End effectors are the last element of a kinematic chain of robotic arms and are used to position tools with a larger degree of freedom than those that are mounted directly to the robot base. The tools can be purely moved to gain a position advantage, such as for inspecting cavities or to measure in proximity of objects, or can be brought into contact with the environment. An example of such latter tools are grippers or conventional tools such as drills, screwdrivers or cutters. In order to extend the robots capabilities, end effectors can e.g. be equipped with multiple tools at once or tool changers.

Effectors as weapon payload: In a military context, effectors might be weapons that generate certain effects. These consist of conventional effectors that generate a kinetic effect on the target, such as small arms, mortars, improvised explosive devices (IEDs), rockets or missiles but also means that emit energy without explosives, such as low or high energy lasers (LEL/HEL), acoustic weapons, microwave guns, or chemical, biological, radiological agents and nuclear weapons.

1.2.4 Communication and Interfaces

System-to-System and inter-system communication: Reliable, flexible and sustainable wireless communication infrastructure (hardware and software) to exchange data between the parts of robotic systems is currently vital for every collaborative task. As an example: Each drone of a drone swarm may need to exchange, among many other data, its position information in order to coordinate flight patterns and formations autonomously. Recognized obstacles or identified objects should be communicated from the internal sensors to the on-board system and distributed from one robotic system to the others. A higher degree of autonomy and rising complexity of intended tasks in multi-robotic systems leads to an increased complexity of internal and external communication within such systems. Besides the fact that communication needs computational power, energy and sophisticated algorithms for processing, it is also – depending on the environment – prone to delay, interruption, misrouting, enemy interference and so on. Especially in search-and-rescue or surveillance scenarios, communication is a necessity – and if it fails, questions about how such systems can coordinate among each other under constraint circumstances are difficult to solve. Current state-of-the-art technology therefore (for these and other reasons) does not include widespread use of multi-robotic autonomous systems and is limited to academic research in controlled environments.

System-to-Human Communication: Most of the mentioned characteristics and problems of system-to-system communication in multi-robotic systems also apply in system-to-human communications. A reliable communication link between the human user(s) and the robotic systems is necessary to maintain whatever degree of control humans desire to have over these systems. Reasonable local and remote control implies a direct or indirect ability to operate the whole system or critical functions within the system in order for the robotic system to complete the intended task(s) correctly. Partial or total communication failure or loss due the wave propagation environment, malfunction, enemy interference (jamming, spoofing or hacking), poor implementation, bad design, interference with the environment or other reasons can lead to a direct and serious impact in real world scenarios. Whereas communication infrastructure is currently almost omnipresent, especially for aerial vehicles, it is still not clear how to deal with unpredictable situations in which humans are unable to remain in contact and interact with the system during ongoing critical operations. As these situations are often unacceptable, proposals to mitigate these problems range from on-board "kill switches" (algorithms that perform security checks and react according to pre-defined decision trees; INTV14) to automatic "return to base" functions.

Human-Machine-Interfaces: Several types of interfaces (hardware) between robotic systems and humans exist today. Examples include simple remote controls with buttons/joysticks, augmented reality devices or omnidirectional treadmills, brain-machine-interfaces, exoskeletons or less futuristic text or graphic terminals. They enable human-to-system interaction. With the help of interfaces, humans can rapidly get information about a system's status and, if required, influence the system through haptic stimuli or brain signals to guide these robotic systems towards specific tasks. Through such interfaces, humans can also be supported in specific tasks. A good example is telemanipulation, which allows for fast decision-making and control; it relies on strong communication in both directions, but can face the same difficulties as system-to-human communication does such as slow and delayed data transfer, imprecise calculation of aggregated data and so forth. This can decrease the ability to interact effectively through these interfaces (NASA 2015, p. TA4-45 – TA4-46). State-of-the-art applications involve unimodal (mostly visual) interfaces, but also multimodal interfaces that combine different sensory information and communication (e.g. brain signals combined with 3-D graphics, tactile sensors etc.). Multimodal interfaces still lack – among other things – precision, speed and compatibility and are still being developed and tested in lab environments.

1.2.5 Energy Supply

All robotic systems need power to operate (or more precisely a converter to convert energy from one form into another) as well as an energy storage to continuously supply the system with power. Depending on the area of application, the size and shape of the system and the budget available – several possibilities with pros and cons are on the market and easily accessible. Possible energy supply systems may include the generation of electrical energy from chemical reactions (electric batteries), from heat flux (thermoelectric generator), from electromagnetic waves/light (via solar cells) or the generation of mechanical energy by combustion engines. In robotics, energy is usually stored or buffered by e.g. electrochemical (battery), chemical (fuel), electrical (in an electrostatic field), mechanical (via springs or rotating masses) or even nuclear means (as with NASAs Mars rover, Curiosity).

The main factors when considering an energy source for robotic systems are capacity, energy density, provided peak power, and durability/robustness, cost and environmental factors, and the shape of the energy source itself (Bohidar et al. 2014). Depending on the need of energy for a specific robotic system, a higher payload capacity and thus a bigger sized robotic system may be required. Thus, the laws of physics can limit the operational ability of a system when it comes to energy supply. For instance, one cannot mount a 2-kg battery on a small-body or lightweight mini drone.

1.2.6 Data Processing and Storage

The automatic collection and processing of data to produce meaningful information (for motion planning and task execution, including image data processing, object recognition, mapping, and communication) can require a significant amount of processing power and memory, especially when handling complex tasks in changing and challenging environments. The required processing units (Central or distributed processing units) can be located on-board or off-board. Within the internal and external communication network, control/computer systems should be implemented in a system architecture that allows for flexibility, compatibility, interoperability, reliability, safety and security. Depending on the given mission(s) and complexity of the system, they may need to process large amount of data. Creating and implementing systems that are able to grow modularly, calculate fault tolerances, reconfigure dynamically, have the ability to coordinate multiple (sub-)units, aggregate data and adjust to on-board needs is a major challenge (Liberatore et al. 2004).

1.2.7 Learning and Artificial Intelligence

We will use the term "Artificial Intelligence" (AI) to refer to technologically implemented correlates to cognitive skills such as intelligence, learning, and other cognitive abilities. The AI field is filled with different contested notions of the definition of intelligence such as "human-level AI", "Artificial General Intelligence", "strong AI", etc. (Sotala & Yampolskiy 2013). A reasonable definition has been provided by the Future of Life institute, which has, *inter alia*, in this context defined the term "intelligence" as "related to statistical and economic notions of rationality – colloquially, the ability to make good decisions, plans, or inferences" (Russell et al. 2015).

Machine Learning: One of the key driving fields in artificial intelligence research is that of machine learning – where computer scientists and mathematicians seek to create algorithms that are, based on data inputs, able to learn on their own and use probability calculations to predict outcomes (e.g. learning from trading patterns to predict the outcome of financial investments). One could define machine learning as "any computer program that improves its performance at some task through experience" (Mitchell 1997, p. 2). There are, however, many different theoretical approaches to machine learning such as symbolic systems (inverse deduction), connectionist systems (such as deep learning), evolutionary algorithms (such as genetic algorithms), Bayesian learning systems, or analogy systems (such as support vector machines).

Generally, machine learning can be divided in the following types of algorithms: Supervised, unsupervised and reinforcement learning algorithms. Supervised algorithms can be applied to classify data based on a training dataset with known classes. With unsupervised learning algorithms, the structure of the data is not known in advance and the algorithm has to come up with a 'classifier' to structure the data. In reinforcement learning, a proposed classification of data can be rewarded if correctly classified, or otherwise punished.

Increasingly discussed are deep learning approaches using neural network models. Artificial neural networks are loosely defined as computers with algorithms that are "modeled on the [biological] brain, and that promised to be better than standard algorithms at dealing with complex real-world situations" (Castelvecchi 2016, p. 21). Whereas deep learning is a hot topic in science and many are excited by possible applications, it seems to come with a hard problem. In deep learning neural networks, learning results are not stored in one specific place, but they rather are "encoded in the strength of multiple connections" (ibid., p. 22) – in the same way it is done presumably in the human brain. Given the fact that such neural networks may have millions of connections and are arranged in many layers, one cannot easily "reverse engineer" the processes in order to understand what the neural network did and how exactly the output results are generated. As an illustrative example, Pierre Baldi asks where exactly a phone number is stored in the human brain and answers "Probably in a bunch of synapses, probably not too far from the other digits" (Baldi 2016; in Castelvecchi 2016). However, "there is no well-defined sequence of bits that encodes the number" (Castelvecchi 2016. p. 22). Therefore, one does not really understand a neural network and one is confronted with some sort of a "black box" (INTV17).

1.3 The Concept of Autonomy in Robotics

Autonomy is a concept with a rich theoretical history. The term "autonomy" or "*auto-nomos*" in the sense of self-governance was originally coined in the time of the ancient Greek city-states, who were aspiring towards independence from the Persian Empire[3]. In the age of Enlightenment, autonomy was seen as the ability for self-governance that grounds moral responsibility. In this way, autonomy became a fundamental concept of (moral) philosophy. It reflects a basic moral and political value, affecting how individuals interact, and the moral and legal rights and responsibilities they are given.

However, when the terms "autonomy" or "autonomous" are used in the context of robotics, they do not refer to those high-level philosophical concepts. In robotics, a purely operational understanding of autonomy is usually used that – according to some authors – may include even simple types of automation. According to an operational definition, a robot is autonomous if it can complete a given task without human intervention (however, human supervision might still be allowed during an autonomous execution of a mission). In this sense, it is possible to ascribe operational autonomy even to simple machines. For instance, a toaster is autonomous insofar as the machine can properly react when a bread slice is warm enough and eject it accordingly. Similarly, the heating system of a house is autonomous insofar as it can sense the temperature in the interior environment and decide to switch the heaters on and off in order to bring the room temperature to the desired level at the desired time. In this simplest form, operational autonomy is equivalent to automation. However, it is a controversial issue, whether automation should be considered as a form of operational autonomy. Russell and Norvig (2014) for instance would limit the use of the term "autonomy" to its more complex forms, like those we describe below in the remaining part of this section. Also, according to van den Vyver et al. (2004), machine autonomy consists of more complex properties: (a) the ability to make independent decisions based upon observations, to do planning, to draw conclusions and to make judgments concerning consequences; (b) the independent completion of tasks, by combining the planning and controlling steps; (c) the ability to learn and eliminate mistakes; and (d) the ability to cooperate with humans and other machines.

[3] αὐτονομία autonomia from αὐτόνομος autonomos from αὐτο- auto- "self" and νόμος nomos, "law" a person who legislates one's own law.

In this report, we consider autonomy in robotics to be a gradual concept that develops along several dimensions.

We will consider systems that score low on those dimensions not as autonomous robotic systems. Therefore, whereas simple automated systems such as land mines surely can pose serious and ethical issues, these simple systems are not in the focus of this report. On the other hand, there might be in the future systems available that score very high on those dimensions and thus would enjoy moral autonomy ("artificial moral agents", see section 4.2.1.). Nevertheless, robots of the present and of the near future are likely to be equipped only with operational autonomy. As robots will be equipped with more and more complex forms of operational autonomy, future developments in autonomous robotic technologies may still have serious ethical implications (they may become "ethical impact agents", see section 4.2.1.).

Operational autonomy may have many different dimensions, depending on, among other things, the component of the system to which it applies (sensing, thinking, acting), and on the different ways in which these capacities are realized, as well as on the different ways in which the other (autonomous) parts of a system (including human operators) interact. The following is a list of the most salient dimensions along which to assess robotic systems autonomy:

- **System autonomy as a measure for the "control distance" from the human operator:** Technological systems have a purpose, i.e., they are built to execute certain tasks relevant for humans. Thus, the "control distance" of the human operator is a first, intuitive measure for system autonomy. Primitive tools such as a hammer only function when *directly operated* by humans and the human is also the provider of the energy needed for functioning. Still today many of the machines we use are under direct human control; i.e. they are turned on and off by the human operator and they are steered by the operator in most of its relevant functions. However, the controlling of relevant functions of machines can be predetermined by the operator, which leads to *automation*. This is one way to increase the control distance. Another element of distancing is that the human operator can *remotely control* systems. This usually affects situational awareness of the operator, who may become dependent on sensory input (from the system itself or from other systems or observers). Yet another element of distancing concerns the *degree of the precision of the orders* given to the system. The less precise the orders have to be, the larger is the control distance between system and operator.

- **System autonomy compartmentalized for the system functions:** Another approach to characterize system autonomy is to look at degrees of autonomy for specific functions of the system. In a report of the Defense Science Board (2016), for instance, the authors distinguish between the functions "sense" (sensors, perception, fusion), "think/decide" (analysis, reasoning, learning), "act" (motion, manipulation) and "team" (human/machine, machine/machine, and information exchange). Here, one determines the degree of autonomy of each of those functions in order to characterize the overall autonomy of the system. Interaction between those autonomous components could lead to emergent behavior and to unwanted risks.

- **System autonomy captured by spatial and temporal time scales:** Another possible measure for system autonomy concerns the spatial range and temporal duration in which a system operates without human intervention. This aspect is mainly captured by the mobility of the robotic system itself. The more and the longer a robot is able to move around the higher the number of different objects and scenarios it may encounter, and the higher the amount and complexity of decisions it must make. Those spatial and temporal measures obviously strongly depend on the complexity of the context in which a system operates: complex, urban environments for example pose different spatial challenges compared with flat, monotonous landscapes; and situations that include fast-changing aspects are harder to handle. Characterization of these environmental factors will lead to a definition of the "operational design domain" of the system.

- **System autonomy captured by task complexity:** In general, one can also characterize the autonomy of a system along the complexity of the task it has to execute. Both a smart toaster and an autonomous vehicle are operationally autonomous. However, to decide autonomously when a slice of bread is toasted usually requires much simpler capacities than autonomously deciding how to steer a vehicle along a busy urban road. One factor that makes a mission more complex is the number and complexity of decisions that the system should make. Imagine a robot designed to pick and move objects in a warehouse; it may drop an object by mistake: deciding what to do and how to do it in order to address this unexpected situation is something more complicated than just repeating a few tasks across a limited range of different conditions (INTV17).

- **System autonomy captured by the complexity and structure of the environment:** Some autonomous systems operate in an environment that, by nature or by design, presents a stable and easily recognizable structure and does not change over time. Other systems operate in environments without a stable structure and/or changing across time. This difference is important for autonomy. As examples of the former, one may think of an industrial robot autonomously repeating the same task in a highly structured factory floor; or, a bit further down the line, a robotic system moving objects within a structured harbor. The more the environment becomes complex and dynamic – one may think, again, of an urban road – the more flexible the capacity for sensing and deciding of the robotic system must be; and the higher the risk the system encounters an object or a scenario that it cannot properly recognize or handle.

- **System autonomy captured by the interaction and communication with other human and non-human agents:** Robots may have to function while interacting with a variety of human owners, instructors, collaborators and end-users, as well as with animals and other robots. Hence, a robot's behavior will be shaped by a variety of other agents, each with its own individual preferences, styles and cognitive capacities. The person owning the robot may not always be present when the robot is in operation, the users instructing and collaborating with the robot ('instructors') may be different from the owner, and the behavior of the robot may affect persons ('end-users') who are neither owner nor collaborator. For instance, in a care-giving context, the institution may be the owner. Professional caregivers may be the instructors as well as collaborators of the robots, and the elderly are end-users experiencing the effects of the robot's actions. An autonomous vehicle would be programmed by a car manufacturer, operated by a private person, while also interacting with and affecting the behavior of other road users. This interaction with other intelligent systems adds a significant layer of complexity to the autonomy of the robotic system. Thus, the degree of autonomy of a system can be related to the number of different types of interactions the system is able to perform. Interaction of systems with autonomous capacities and human operators are particularly relevant with respect to ethical issues. For example, one fully expects an autonomous system to provide inputs in addition to human inputs (thereby taking advantage of system autonomy). But one should be cautious about a system that pro-vides inputs contrary to human inputs. For example, consider the 2008 US B-2 stealth bomber crash. The aircraft crashed on initial takeoff injuring both pilots and destroying the $1.2B aircraft when "the computer … told the aircraft it was going into a nosedive, when the pilots were actually in the process of lifting the craft off the ground. The computer ordered the B-2's nose to pitch up 30 degrees" (Burgess 2008). The pilots (a) were unaware of the computer's alternative inputs to the control system and (b) had no means of taking control of the aircraft back from the flight computer.

- **System autonomy captured by the autarchy of the system:** Robots need energy to function and usually require maintenance in case of, e.g., damage or wear. However, the less a system is dependent on external energy input (e.g., because it has solar cells to generate own electrical energy) or the greater its ability to "self-repair," the more independent can the system function from human up keeping, i.e. the autarchy of the system is increased.

- **System autonomy captured by capacity for self-organization and collective agency**: A specific dimension of autonomy is that displayed in multi-robot systems such as swarm robots. Swarm technologies aim at creating groups of robots that operate without relying on any external infrastructure or on any form of centralized control (Dorigo et al. 2014). In a robot swarm, the collective behavior of the robots results from local interactions between the robots and between the robots and the environment in which they act. A robot swarm is thus a self-organizing multi-robot system whose collective behavior emerges from the interactions of each individual robot with its peers and with the environment. This form of autonomy may produce a more robust and flexible system. However, it may increase risks due to decreased control and predictability on the part of programmers (emergent behavior).

More dimensions of autonomy could be described. Moreover, dimensions often interact (e.g., task complexity with environment complexity) and are usually hard to quantify; increases in one or more dimensions may also add complexity and give rise to moral problems

1.4 Major Trends and Challenges in Autonomous Robotic Systems

New developments in robotics concern both the types of robots themselves as well as human robot-interaction: With respect to new types of autonomous robotic systems, the following trends are remarkable:

- **Self-driving vehicles:** This is an area where serious progress has been made in the past decades. In 2004 not a single robotic car managed to complete a two-hundred kilometer course in the Mojave Desert, but only a year later five out of 23 robotic cars managed to complete the full course in the required time. In 2012, Google created its first self-driving car. Whereas more and more progress may be expected in the years to come, many outstanding researchers in the field are not optimistic that these vehicles will be able to safely move any time soon in a real, dynamic, not-completely structured environment like urban traffic (Shladover 2016, Cummings 2016). As John Leonard puts it, this seems to require not a specialized but a general reasoning capability, which is difficult to achieve in the short run (De Crook et al. 2016, pp. 57-60).

- **Rescue operation robots:** Robin Murphy, an expert in rescue robots at Texas A&M University, claims that state of the art rescue robots primarily allow human responders to see at distance. She claims that thanks to underwater robots, the survey of Japan's coastline infrastructure after the 2011's tsunami was completed six months earlier than would have been possible with human divers alone (De Crook et al. 2016, pp. 50-54). According to Murphy, the use of robots in rescue operation becomes more and more common. The first time they were deployed was after 9/11 in New York (Murphy 2004); since then they have been used at least in 47 disasters in 15 countries. Rescue robots include unmanned ground vehicles (UGVs), unmanned aerial vehicles (UAVs), and unmanned marine vehicles (UMV). The potential for gathering data is massive: in the 2015's Blanco River flood in Texas, a single 20-minute UAV-flight produced over 800 high-resolution images. According to Murphy, future rescue robots may be snake-like and will be able to dig into rubble to reach people trapped below (ibid.).

- **Healthcare robots**: Most prominent and widespread examples of robots in healthcare are surgical robots like the Zeus™ Telesurgical System and the daVinci™ Surgical System; however, more and more research is being done in robots to be included in daily care activities of persons like feeding and bathing. These robots have high potential in enhancing the autonomy of the persons by allowing them to take care of themselves without any human help, but they also raise the concerns of lowering the quality of care (dehumanization), and bringing unwanted risks for the safety, dignity, and privacy of patients (van Wynsberghe 2015).

- **Social robots:** Whereas research in so-called social robots has developed greatly in the past decades, applications are still quite limited. Robots have proven to be able to assist professionals in specific tasks, for instance portions of healthcare practices (lifting patients), rehabilitation activities, and support of therapeutic programs for children and patients with mental disorders. Robots may be able to interact with workers on the factory floor, but the robots of today are far from being able to interact smoothly and safely with untrained humans in an unstructured environment where a complex and possibly delicate task has to be performed, as is the case in rescue operations. While a great amount of research is being done in this field, including affecting computing, and techniques for bi-directional, non-verbal communication between humans and robots, whether there will be a breakthrough in the next 10-15 years in this area as there has been in sensing, data processing and autonomous vehicles remains an open question (INTV17).

- **Biologically inspired robotics**: The main idea behind this area of research is to look at biology for inspiration based on the consideration that the behavior of animals is extremely flexible and robust in the face of environmental contingencies. The hope is that adopting some of the design principles of animals will endow robots with similar flexibility and robustness (Beer 2009). Examples include legged locomotion, climbing robots, and robot swarms (see section 1.2. above).

- **Autonomous UAVs:** According to Chris Verhoeven (INTV18), from a *purely technical* point of view, the most efficient way to make the most of the sensing and data processing technologies available now and in the near future would be having tiny "informants" in the form of small drones constantly flying autonomously around and collecting information. Particularly valuable could be their capacity to provide real time information about a sensitive or potentially dangerous situation, thus, e.g., leading to the possibility of predicting disasters before they occur. Such a system would at least help to prevent the worst consequences of a disaster or reduce its impacts as well as providing real time information to manage such situations when they do occur. This outcome is only possible if the system is already deployed in an area and would not be as useful if it were only sent after the fact. The same applies to surveillance and crime prevention. According to Robert Babuska (INTV17), in the short run an important distinction is to be made between the use of autonomous drones in indoor or otherwise structured or controlled environments (e.g.

big tanks, harbors), and their use in outdoor, unstructured environments. The former use is likely to be feasible in the near future, whereas additional issues may arise with the latter use, especially in interaction with humans.

- **Nanorobotics**: Nanorobotics is an emerging area, still largely theoretical, based on nanoscience and nanotechnology, which covers a wide spectrum of enabling technologies such as micro/nano-sensors and actuators, power supply, manipulation, control, and embedded processing. It is considered particularly promising in relation to bio-medical applications, such as medical image processing in wireless capsule endoscopy (Guo 2013).

With respect to human-robot-interaction, the following trends are remarkable:

- **Wearable robots:** Unlike what happens with conventional robots, wearable robots are designed to work via direct physical interaction between the robot and the human operator (they are literally worn by the operator). The interaction with humans include physical and cognitive aspects: the control of functions is typically shared by human and machine. Wearable robots can be used either to augment, train or supplement motor functions or to replace them completely. Wearable robots operate alongside human limbs, as is the case in orthotic robots, exoskeletons or robotic suits. Wearable robots have a high potential for applications in medical, industrial, and consumer domains, such as neuro-rehabilitation, worker support, or general augmentation. However, exoskeletons for enhancing rather than just restoring human capacities have been difficult to realize so far, at least for military purposes. The exoskeletons developed so far, according to Hugh Herr from MIT, are too bulky and tend to fight the natural rhythms of the body, which turns them into exercise machines rather than enhancers (Cornwall 2015).

- **Brain-machine interfaces**: Brain-computer interfaces (BCI) establish a direct communication pathway that allows users to control an external computer device exclusively with brain activity, bypassing the peripheral nervous and muscle systems. BCIs were originally developed as a therapeutic or assistive technology for neurological patients; they are used to repair, assist or augment cognitive or sensory-motor functions in patients with cognitive or sensory-motor impairments (Allison et al. 2007; Vallabhaneni et al. 2005). BCI based motor prostheses have successfully been trialed in animal models and patients to enable direct brain control on artificial limbs, wheelchairs and other devices (Fetz 2015). BCI applications have become available also to the public. Especially in combination with wireless technology, BCI has a high potential insofar as it might allow for the direct control of animal and, in principle, human brains and behavior at distance (Borton et al. 2013). It certainly also raises serious ethical concerns, namely because of the risk of hacking (Ienca & Haselager 2016) and violations of human autonomy, as the brain enables human self-regulation and self-determination.

- **Robot testing and human training with autonomous robotic systems**: In addition to technical challenges, developments in autonomous robotics also bring challenges from a broader socio-technical perspective. There are two notable general challenges. First, in order to improve the safety of autonomous, interactive systems reliable tests of the interaction between complex systems and real people in a real environment are necessary. However, tests with real people in the real world are morally and legally problematic until technology has proven to be safe enough. This may lead to a stalemate in the progress of innovation. One proposal that has been advanced is that of creating “special zones” for testing robotic technologies under the cover of special legislation (Weng et al. 2015). Second, the challenge arises of providing appropriate training to professionals and laypeople who may have to use or interact with autonomous robotic systems. A smooth interaction between human and non-human agents requires the best possible training on both sides.

2 Security

This chapter outlines the current and future use of systems with some degree of operational autonomy in the security sector. It briefly defines what is meant by "security sector", presents types of autonomous systems currently in use in this sector, identifies difficulties in using such systems in military command and control structures using case studies, and provides an outlook on future developments.

2.1 Defining "Security" and "Security Sector"

In this chapter, the discussion of the application of autonomous systems is limited to the security sector. This report uses the term "security sector" to refer to the large-scale economic markets interested in autonomous systems capable of coercing behaviors or protecting/rescuing persons. The most common actors in the security sector are governmental. They usually involve military, law enforcement, or emergency response and rescue organizations. However, it is possible for private companies to act as contractors for government agencies in this field and therefore for commercial organizations to participate in the security sector. Because the scope of this section is limited to the security sector, the use of the term "security" is limited to refer only to the security provided by security sector organizations (government agencies and private military/security companies).

This report defines "security" in terms of "the security sector" and not the other way around and this will yield some important implications. For example, financial organizations may employ semi-autonomous or autonomous systems in the cyber domain to secure assets against hacking. Nevertheless, these systems are employed in the commercial sector rather than the security sector. Therefore, they fall outside the scope of this section. A state's Ministry of Defense might use the very same kinds of autonomous systems in similar ways to protect state secrets. Because it is employed by a security sector organization, however, this kind of security would fall within our definition of "security." Thus, the primary subcategories with which this section is concerned are the application of operationally autonomous systems in military operations, law enforcement operations, and emergency response and rescue operations.

Classifying autonomous systems within these three operational categories (military, law enforcement, and emergency response and rescue) is difficult because the requirements vary significantly from one category to the next. For example, domestic law enforcement agencies are interested in keeping peace and using violence only as a last resort. There is a requirement, for example, for law enforcement officers to have alternatives to firearms. The United Nations Congress on the Prevention of Crime and the Treatment of Offenders suggests that "governments and law enforcement agencies should develop a range of means as broad as possible" to include "the development of non-lethal incapacitating weapons" (United Nations 1990). Therefore, the weapons used by domestic law enforcement agencies must cover the whole range demanded by the escalation of force, including both lethal and non-lethal means. As a result, the weapons that are intended to be non-lethal (e.g., Tasers) are important *because* they are (intended to be) non-lethal. Likewise, those intended to be lethal (e.g., firearms) are important for their lethality. Muddying the waters, however, is the empirical fact that so-called non-lethal weapons have in fact killed people and so-called lethal weapons have in fact caused non-fatal injuries (Haar & Iacopino 2016). In the military context, the intent is, broadly speaking, to target the enemy's war-fighting capability. Because there will be harms associated with such actions, ethical and legal frameworks that seek to govern military weapons do so with respect only to the harm such weapons cause without distinguishing between weapons intended to be lethal and those intended to be non-lethal. There is no standing apparatus for legally evaluating so-called non-lethal military weapons. The International Committee of the Red Cross (ICRC) claims that the Article 36 requirement for legal reviews of new weapons applies to weapons broadly, "be they anti-personnel or anti-materiel, 'lethal,' 'non-lethal,' or 'less than lethal.'" (ICRC 2006, p. 9). Nevertheless, there is no separate set of criteria for evaluating the legality of weapons intended to be non-lethal or less than lethal other than that which applies to weapons that are intended to be lethal, grounded in the expected harms resulting from their design and their intended and expected use.

2.2 Types of Autonomous Systems in the Security Sector

The vast differences in requirements for autonomous systems across the three operational categories introduced above makes the task of classifying such systems difficult. A look at current technologies with different degrees of autonomy in each sector may help to clarify the difficulty and take initial steps toward solving it.

2.2.1 Emergency Response and Rescue

There have been significant developments in the application of robotics to search and rescue operations in the recent past (see also section 1.4.). Organizations such as the Center for Robot-Assisted Search and Rescue (CRASAR) have devoted considerable resources, and made some significant strides, in adapting robotics technology to search and rescue operations. Thus far, however, the most important added value of these systems is their ability to operate in situations that are unsafe for humans (as in the Fukushima Daiichi nuclear accident)[4] or in spaces so small that humans cannot fit (as in the World Trade Center rescue operations that followed the 9/11 attacks).[5] However, the majority of robotic applications in the search and rescue sphere have been remotely controlled (via radio frequency communications or a wire tether) and thus, the dimension of autonomy such systems employ has been generally limited to "control distance" (see section 1.3). The only CRASAR robots capable of autonomous operations (in this case, image collection) in real-life emergency scenarios are two aerial systems called Rita and Virginia that "can be preprogrammed to collect images autonomously."[6] Such systems employ degrees of autonomy in six of the seven dimensions described in section 1.3 (Control Distance, Compartmentalized Functions, Spatial and Temporal Scales, Task Complexity, Environmental Complexity and Human-non-human interaction) but not self-organization and collective agency.

The small size and maneuverability of the robots involved in post-9/11 rescue operations allowed them to penetrate areas of the rubble that people were unable to reach. This capability yielded some unexpected results. The US military quickly recognized the value of such systems in the search for Osama bin Laden and other al Qaeda leaders in the mountains of Afghanistan and began to pursue similar technologies for such a purpose (INTV20). This is just one example of a technology developed for one of the three operational categories introduced above, and subsequently transferred to a different operational category. One lens through which to view this phenomenon is the admission that, in addition to systems intended for dual-use (systems that were designed to be used in more than one operational category) there may be unforeseen and unintended transfers as well.

[4] As the iRobot Packbot 510 did. CRASAR, "Heads Up, SUGV!," http://crasar.org/robot-petting-zoo/heads-up-sugv/.

[5] As the Inuktun microVGTV – named Bujold – did. CRASAR, "Bujold's Rock Climbing Wall," http://crasar.org/robot-petting-zoo/bujolds-rock-climbing-wall/. See also Snyder (2001).

[6] CRASAR, "Rita and Virginia, the Sywriters," http://crasar.org/robot-petting-zoo/rita-and-virginia-the-skywriters/.

2.2.2 Law Enforcement

Domestic law enforcement organizations already employ systems with some limited autonomous capability that can determine automobile speed, assess that real-world speed against the posted speed limit, and in the case of a violation, record the vehicle's license plate number. These systems are often tied into government data centers or networks such that the citation is automatically printed and sent by mail to the offender (DC Metropolitan Police Department).

There have also been kinetic uses of robotics technology in law enforcement applications. In a contentious turn of events, the Dallas Police Department (US) recently appropriated military robotics technology to a domestic law enforcement application. The Remotec Androx Mark V A-1 delivered a 1-pound C4 payload to target a sniper that had been shooting at police officers in July 2016 (Sidner & Simon 2016; Karimi et al. 2016). This system employs very little autonomy and, like many of the rescue robots mentioned above, the only dimension of autonomy it employs is "control distance" (see section 1.3). Nevertheless, this system that allows increased "control distance" was employed in this way for the first time (Peterson 2016). Though it represents a true first, and therefore an important step toward further technological automation in US domestic police activity, it does not appear to have violated US law (Roberts 2016), though the use of explosives in law enforcement operations is a controversial issue from an international human rights law perspective.

There are at least two important issues at stake in this example. First, there is the migration of equipment that is traditionally and originally associated with the military (the robot itself, and the C4 explosives) to domestic law enforcement use. This transfer is analogous to the transition of rescue robot technology to military high-value target search applications especially in countries such as the United States (see section 2.2.1). Much has been written on the rapid growth of military-style equipment among domestic law enforcement agencies and such growth has been ongoing for decades (Kraska & Cubellis 1997). A thorough account of this phenomenon falls outside the scope of this report. Here we need only notice that though the human rights law that governs law enforcement activity and the international humanitarian law that governs military action stand quite apart, there is an increasing cross flow of equipment from the military to the law enforcement context and this may generate some causes for ethical and legal concerns.

Second, though the Dallas police robot is better characterized as remotely operated than autonomous, it nevertheless was used to hold the sniper at risk while reducing the risk to the police officers involved. Should the Dallas example become a precedent and operations of this kind become a trend, it might prove to be a trend that many find uncomfortable. Though in the military context, increasing risk to one's enemies while decreasing risk to one's own forces (provided there is no parasitic increase in risk to non-combatants) is not merely permissible, but it is often an intended purpose of military operations. In the domestic law enforcement case, however, the normative purpose of such operations, according to the UN Congress on the Prevention of Crime is to "apply non-violent means before resorting to the use of force and firearms" (United Nations 1990). One wonders whether armed robots will naturally lend themselves to unnecessarily rapid (and therefore perhaps unjustified) escalations of force that unnecessarily increases risk to civilians thereby imposing a kind of moral hazard.[7]

[7] For comparison, find a well-thought out application of this moral hazard in the military context in Kaag & Kreps (2014).

Another less controversial application of autonomous systems to law enforcement operations is a commercially available system that identifies the point of origin of a discharged firearm without required inputs from police officers or witnesses. Where installed, "ShotSpotter Flex" measures the latency of the shot's sound in multiple microphones at varying distances and directions. It then autonomously triangulates the point of origin and transmits that location to police vehicles in the vicinity (ShotSpotter). As was the case with some of the rescue surveillance systems listed above, the autonomy of this system can be measured in terms of six of the seven autonomy dimensions in section 1.3 (Control Distance, Compartmentalized Functions, Spatial and Temporal Scales, Task Complexity, Environmental Complexity and Human-non-human interaction) but not self-organization and collective agency.

2.2.3 Military

Examples of autonomous systems in the military context include the US Air Force and Navy RQ-4 Global Hawk, the US Navy's Mk-15 Phalanx Close in Weapons System, Israel's Iron Dome missile defense system, South Korea's SGR-1 defense weapon, and the UK's Taranis technology demonstrator. Though the Global Hawk can be controlled real-time, it is autonomous in the sense that it is capable of flying preprogrammed, Intelligence, Surveillance, Reconnaissance (ISR) missions from start to finish (Drezner & Leonard 2002). Israel's Iron Dome missile defense system identifies inbound rockets and missiles, determines whether those projectiles are directed toward residential areas, and if so, automatically fires a missile to intercept the inbound projectile (Berman 2012). Though a human may be able to interrupt or "abort" the engagement, human action is not required for the engagement to commence. The US Navy's Mk-15 Phalanx Close In Weapons System, though used primarily for fleet defense, has also been deployed in a land-based force protection role. It is capable of autonomously targeting aerial threats that meet a predetermined set of flight profile conditions. A subsequent iteration of the system includes an optional setting that requires human crewmembers to visually confirm targets prior to engagement (Raytheon Corporation). South Korea's SGR-1 is deployed to the demilitarized zone between North and South Korea and, though it is capable of targeting and engaging dismounted personnel autonomously, designers have included the requirement that a human operator consent to weapons employment prior to engaging enemy soldiers (Prigg 2014). Finally, the UK's technology demonstrator "Taranis" is often listed as an autonomous weapons system and it can conduct entire missions autonomously. The degree of autonomy with respect to targeting and weapons employment, however, is a closely held secret (Del Prado 2015).

2.3 Status of Autonomous Capacities in Military Command & Control Structures

One should recognize that while the current level of technological sophistication and algorithmic control may be quite a recent development, the fact of autonomy is not. Conceptually, these systems function based on a series of if/then determinations to reach the "launch" or "engage" decision. For example, in the Iron Dome case, (1) *if* a projectile is detected, (2) *and if* it has the parabolic flight path of an inbound rocket or mortar, (3) *and if* the projectile is pointed toward a residential area, *then* a missile is launched in response. Each of these conditions is evaluated technologically without human input. The same kind of autonomy, though admittedly lacking in technological and algorithmic sophistication, has been present in land mines for over 100 years. A land mine is – following our considerations regarding autonomy in section 1.3. – rather an automated weapon with an abysmal record of meeting the discrimination and proportionality requirements of international humanitarian law and the just war tradition. The fact that "operational autonomy" is a category that admits cutting edge and future technologies as well as centuries old technologies is thus another good reason to consider (operational) autonomy on a sliding scale and to define such systems based not on the false autonomous/not autonomous dichotomy, but on a graduated spectrum of less to more autonomous according to the dimensions listed in section 1.3.

In addition to stand-alone autonomous systems, a number of US military leaders see human-machine teaming as the next immediate step in autonomous systems development. Applications in this sphere would range from ISR operations to logistical support to kinetic combat operations. For example, Mr. Matt Donohue, Director of the US Army's Ground Maneuver Technology Portfolio suspects that autonomous systems will develop "layer-by-layer" and that the next such development will allow for a convoy "leader-follower" capability (McNally 2014) allowing a manned vehicle to lead a convoy of unmanned follower vehicles (Judson 2016). Likewise, Dr. Greg Zacharias, the US Air Force's Chief Scientist, has said that "truly unmanned" air vehicles will be partnered with a manned aircraft flight lead as "loyal wingman" within a decade (Malenic 2016). The US Air Force *Small Unmanned Aircraft Systems Flight Plan 2016-2036* suggests that ISR functions such as onboard processing, exploitation, and dissemination (PED) will be 90% autonomous within the next twenty years, that swarming technologies will be half human-controlled and half autonomously-controlled, but that targeting operations against human targets will still remain entirely under the control of a human decision-maker (ISR 2016). Bob Work, the US Deputy Secretary of Defense, has summarized these views saying that "the way we go after human-machine collaboration is allowing the machine to help humans make better decisions faster" (Pellerin 2015).

2.3.1 Case Studies

The difficulty that naturally arises in classifying such systems is grounded in a number of intersecting concepts. The following examples will be assessed in terms of the dimensions of autonomy introduced in section 1.3: (1) control distance, (2) compartmentalized functions, (3) spatial and temporal scales, (4) task complexity, (5) environmental complexity, (6) human and non-human interaction, and (7) self-organization and collective agency. Consider a few notional examples.

Emergency Medicine System on the Battlefield: As a first example, imagine an autonomous medical system tasked with conducting battlefield triage of friendly combatants, enemy combatants, and non-combatants wounded in the fighting. Two *prima facie* problems arise: First, what if the system malfunctions? Or, put a better way, what if the system encounters an operational reality that its designers did not foresee? There may be a case in which a human doctor would have chosen to treat person A ahead of person B for a host of complex reasons and that the system's capabilities are insufficient to manage that complexity. The end result could be a catastrophic decision that causes someone to die who would otherwise not die. In this medical, non-combat application, there are nevertheless life-and-death consequences for failure. The second concern may present itself even if the system is performing as intended. Here, suppose the system makes determinations based on the severity of wounds and likelihood of revival according to the same variables on which a human doctor would decide. There may be ethical concerns over sacrificing one human life for another at the direction of a robot, even if the structure and the outcome of the decision-making process is identical to the one that would have been made by the human. In this case, we see the following dimensions of autonomy at play: Even if the system is adequately programmed and prepared for its (2) compartmentalized functions and (5) rather extreme environmental complexity, (1) the level of human interaction combined with (4) task complexity yields a potentially ethically dubious result.

Drone Surveillance: In a second case, consider the real-world example of a US Air Force MQ-1 Predator pilot tasked with observing and collecting patterns of life on a high value individual (HVI) in Afghanistan in 2012. Each day, the HVI walks to a neighboring field where children often play, takes a child and places that child on the back of his motorbike before proceeding to conduct his duties as an operational leader of al Qaeda in Afghanistan. The HVI is obviously using the child as a human shield, attempting to prevent the US from striking him. After three weeks of observation, the pilot in command of the Predator is told that the ground force commander would like to strike the individual today. The pilot asks a number of the officers and soldiers involved why it has to be today – why not yesterday and why not tomorrow? He is unable to get a satisfactory answer from the ground force that is directing the strike. He is told only that the proportionality, discrimination, and necessity requirements of the international laws of war have all been met. The pilot, on his authority as an officer and the commander of the aircraft, refuses to take the shot, leaving both the child and the HVI alive. Imagine an attempt at automating this targeting process in some future weapons system. It seems that the only standards available with which to equip the autonomous targeting system to make its determinations are those provided by the Just War Tradition and International Humanitarian Law. But these demands of proportionality, discrimination, and necessity *were* met in this case. Therefore, where a human resisted and chose not to engage on ethical grounds or grounds of conscience, it is difficult to imagine a system that is designed in such a way that it would likewise choose not to engage. Here, the concepts at play include (1) the control distance between the human operator (and his or her conscience) and the system, (2) the compartmentalized functions (in that only the objective international humanitarian law considerations, and not considerations of ethics and morality as such, can be plausibly delegated to the system, and (6) human interaction, given that the ethical content of the scenario is grounded largely in the presence of a child. It is unclear whether the user would *want* the system to violate the standing guidance (i.e., proportionality, discrimination, necessity) for some other, perhaps poorly defined and likely deontological moral intuition.

Rescue Robot: In a third case, consider a notional autonomous robot capable of digging through enormous pieces of rubble to rescue trapped persons. The intentionality of such a system seems undeniably good (it is not designed to harm and it is designed to help innocent people in dire need of help). Nevertheless, in order for the system to move heavy pieces of concrete it would have to be quite large, and therefore probably dangerous to unprotected humans. Imagine a case in which such a system is employed in an area that is dangerous for human rescue personnel. Consider a notional example like the 2001 New York World Trade Center such that the first tower had fallen but the second had not yet fallen. Rescue workers responding to the first tower's collapse would be subjected to the very high risk of the second tower's collapse. Sending in the autonomous rescue robot would reduce the risk to responders. Such a system, though it is capable of removing large pieces of concrete, is also capable of crushing the very people it was designed to help. Here, once again, there are multiple, intersecting concepts at work. These are, most importantly, (6) the impacts of the system's interaction with human victims, but also the (4) task and (5) environmental complexity in association with that human interaction. Finally, in order to provide any safety to human responders, the system must operate at some (1) control distance, making it more difficult for human operators to engage a "kill switch" to shut the system down if something were to go wrong.

Wide Area Motion Imagery System: Finally, consider as a forth case a military wide area motion imagery (WAMI) intelligence, surveillance, and reconnaissance (ISR) system. Such a system (already being pursued by the US Air Force) would provide imagery coverage of more than 100 square kilometers. Current bandwidth limitations are such that the system is unable to transmit the entire coverage area to the ground-based intelligence personnel who provide processing, exploitation, and dissemination (PED). So the system transmits only "video chips" and "subviews" (Menthe et al. 2012). Suppose the determination of which video chips and subviews to distribute to ground-based intelligence personnel were automated such that the system would decide, based on pre-programmed factors, which video would be seen by the human user in real-time. Such a system might prove a valuable decision-aid for military commanders. Such a system might be programmed to maintain awareness of known enemy vehicles, then identify times and places in which the vehicles are co-located and transmit that "subview" to the ground crews, allowing for higher resolution PED. While potentially valuable, such a capability comes with some opportunity cost. While the system autonomously directs the intelligence personnel (and by extension, the commander) to a particular area of interest, given the bandwidth limitations, it necessarily does so at the expense of other areas. It is easy to imagine, as was the case with previous notional examples, that the system might perform exactly how it was intended, and yet miss an important element because the importance of that particular element was unforeseen by the designers. In such a situation, the machine autonomously directs the commander's attention toward something and therefore also away from something else. This is precisely what happened to human operators (without contributions from artificial intelligence) in the inadvertent 2015 AC-130 strike against a Doctors Without Borders hospital in Kunduz, Afghanistan. The aircrew focused so heavily on the building under their crosshairs – and as a result failed to look at other possible buildings – and convinced themselves that the description they received of the enemy compound matched the hospital (Hickman, 2015). In the Kunduz case, human error led to task saturation, confirmation bias, and channelization. It is not difficult to imagine that an ISR decision aid might, even while perfectly performing its pre-programmed code, result in the same kind of error (Caliskan-Islam et al. 2016). This difficulty can be viewed through the autonomy dimensions of (2) compartmentalized functions, (3) special and temporal scales (e.g., for how long is the system permitted to operate autonomously before its software must be updated to reflect new operational realities and observations), (5) environmental complexity, owing to the difficulty in determining *a priori* which variables will be tactically important, and (7) the dangers of a learning system identifying new important variables incorrectly without human interaction and correction.

These four examples are by no means intended to be exhaustive nor conclusive. They are instead intended to be instructive. The real-world application of autonomous systems, regardless of the context, generates difficult ethical and legal challenges.

2.4 Outlook of Likely Developments

Autonomous technologies will undoubtedly continue to develop. Insofar as the military and law enforcement contexts are concerned, however, there are three ways in which the deployment of these systems will be limited; or perhaps more helpfully, three perspectives from which one can view limitations on forthcoming developments.

2.4.1 Technological Limitations

First, the development of autonomous systems will be limited by technological restrictions. Experts disagree as to how far the technology will develop in the next ten to fifteen years. Varying conceptions include the following important distinctions. Some recognize the difference between being ethical and behaving ethically. One expert claims that the technology is "not anywhere close" to being able to produce moral reasoning (INTV04). Another expert suggests that "an absolute minimum criterion for ethically justified killing is the ability to grasp the moral context of killing" (INTV03). If such a "grasp" or understanding on the part of the system is admitted as a prerequisite (a claim that falls outside the autonomy dimensions described in section 1.3) for the ethical use of lethal autonomous systems, then the technological ability to design and program moral reasoning stands as a significant barrier to ethically justifiable lethal autonomous systems.

Another strictly technological concern regards transparency. On the one hand, in order to produce the kinds of military effects leaders seem to be seeking, the advertised systems must incorporate machine learning. But, according to one expert, "machine learning systems are by nature black boxes – we only see the output" (INTV02). Another suggests that the transparency of the learning system depends upon whether it is a statistical learning system (that will admit of some transparency) or a neural net (that will not) (see section 1.3). The fact that such systems can learn provides the capability in question, but the fact that they can learn implies that we may be unable, *ex post facto*, to discover the reasoning for a particular decision. As a result, transparency as to why a system capable of machine learning acted in the way it did may be either inherently impossible or, at the very least, extremely difficult.

2.4.2 Legal Limitations

The second kind of limitation is grounded in legal requirements. The experts universally agree that technological systems must conform to standing international humanitarian law (IHL) in the military context and human rights law in the law enforcement context. One particularly salient IHL requirement is that military actions only be conducted when the military value of the target exceeds the magnitude of the expected collateral damage (ICRC Customary; IHL Rule 1.4). Though proportionality considerations carried out by human military commanders may often be portrayed as a "numbers game," such considerations are in fact not at all simple for humans and therefore are very difficult to automate. There is no consensus among the experts interviewed as to whether the technological developments over the next ten to fifteen years will be able to satisfactorily meet these requirements.

IHL also requires that soldiers discriminate between combatants and non-combatants (ICRC Customary; IHL Rule 1). There is some consensus among experts that the difficulty of the discrimination problem is dependent upon the particular context in question and perhaps even the domain of war in question. It is likely the case that in the next ten to fifteen years, autonomous systems will still be unable to distinguish between the insurgent (combatant) carrying an AK-47 and the farmer (non-combatant) carrying an AK-47 (see Roff, 2014). Nevertheless, they may be able to distinguish between combatant and noncombatant underwater vessels, aircraft, or spacecraft.[8] There is an additional problem with algorithmic approaches to discrimination. Such approaches would likely require systems to identify combatant targets based on *past* behaviors. One result might be that even if an enemy combatant lays down his weapons, the system would be unable to recognize that particular action, and therefore the change in status (INTV09), violating the IHL mandate to accept surrendering soldiers as prisoners of war rather than as combatants (ICRC Customary; IHL Rule 47). Thus, technological developments may make autonomous targeting systems plausible in some domains and contexts but not in others.

An additional legal concern surrounds the function creep (or mission creep) associated with developing technologies. For example, the Dallas Police Department's June 2016 use of a military-style robot carrying a C4 explosive (Sidner & Simon 2016) to target a sniper was the first law enforcement use of a robotic system in this way (Sidner & Simon 2016; Karimi et al. 2016). According to one expert (INTV21), one

[8] We have included spacecraft because there is a conceivable intercontinental ballistic missile (ICBM) exchange in which a defensive missile defense system must distinguish between a hostile ballistic missile in the exoatmospheric phase of flight and a neutral cell phone provider's satellite.

standard by which such an action might be evaluated is the list of equipment the department issues as standard. The Dallas Police Department (and many other US police departments) regularly issues explosives as a means of intentionally detonating explosive material. It probably had the robot for the same reason: for defusing explosives while keeping police officers at a safe distance. So the use of the explosive carrying robot may have been legally justified because the Police Department standardly carries robots and explosives even though this particular application of robots and explosives was both unconventional in the combination of the two elements and unconventional in that the target was an active shooter (rather than an explosive device). The foreseeable legal problem in the law enforcement category is that a means is justifiable if one already has it, but one will only have it if it is justifiable. Thus, incremental changes in the use of such weapons may generate incremental changes in the normative legal standards for their use.

2.4.3 Operational Limitations

The final way of looking at these limitations is from the military or law enforcement commander's perspective. It may be the case that a system that is deemed legal and that is capable of acting autonomously may nevertheless be unable to achieve the commander's intent. We might return to the case of distinguishing between the AK-47 clad Afghan farmer and the AK-47 clad Afghan insurgent. While failing to discriminate between these two generates a legal and ethical problem, it also generates an operational and strategic problem. If the commander intends to win a counterinsurgency war, for example, then winning "hearts and minds" will be critical to that effort (United States Army 2006). As a result, failing to distinguish between combatant and non-combatant is not only a moral and legal failure, but an unwise operational decision. This problem may, under some circumstances, generate peculiarly operational issues (that is, without generating legal and ethical problems), but in the general case there will likely be significant overlap between the ethical and legal requirements and the commander's operational requirements.

This section warrants a brief note about dual use systems. It is difficult to discuss "search and rescue technologies", because any system designed for search and rescue could be likewise used for weaponized military applications (as demonstrated above in the 9/11 search and rescue robots case, see section 2.2.1). There is also an increasing trend in the modularity of such systems. A state or individual might purchase a drone for the purposes of intelligence, surveillance, and reconnaissance (ISR) operations and then modify the drone to carry a weapon (as ISIS has done in Syria; Gibbons-Neff 2016). Because some systems are designed with this modularity in view, it is difficult to define the system on the whole. For example, Insitu's ScanEagle aircraft widely used for ISR operations around the world is capable of carrying 140 after market payloads offered by more than 60 manufacturers. Some of these add lethal capabilities to what would otherwise be an ISR system. Currently, according to interviewed experts, the Swiss government's legal review of proposed weapons systems (and similar reviews of other states including the US) require that the whole system be evaluated. Notionally, if the Swiss government were to legally review and purchased the ScanEagle at a time when no lethal payloads were available and subsequently, such payloads were made available, the Swiss government would have to conduct a new review of the entire system, including new payloads prior to buying any new payloads. This practice seems wise and will likely continue. We mention it here only to suggest that as this trend in modularity continues, such reviews will likely increase in importance, but will also likely become more time-consuming and cumbersome.

3 Law

This chapter gives an overview of the law governing autonomous weapons systems.[9] Given that several of the substantive legal issues have already been discussed in previous sections (notably sub-sections 2.3.1 and 2.5.2) this section strongly focuses on the legal processes currently under way and the genesis of the debate on the international plane; it provides a sketch of the relevant actors and initiatives (section 3.1). The section then only briefly highlights the main issues arising under both the law of armed conflict and the law applicable in peacetime (section 3.2), while linking back to the legal issues discussed in the previous section. It concludes with a short outlook on likely developments (section 3.3).

3.1 Actors and Initiatives on the International Plane

The international legal debate on autonomous weapons systems takes place mostly in Geneva, within the forum provided by the Certain Conventional Weapons Convention (CCW)[10]. The process was broadly set into motion in 2012, although the academic discussion on ethical and legal aspects of autonomous weapon systems started earlier (e.g., Sparrow 2007; Singer 2009; see also the International Committee for Robot Arms Control, founded in 2009 by Juergen Altmann, Peter Asaro, Noel Sharkey, and Rob Sparrow, which has been driving the discussion forward. One document that was instrumental in getting it moving was a report by Human Rights Watch (Docherty et al. 2012), written together with the Harvard International Human Rights Clinic and entitled "Losing Humanity: The Case against Killer Robots"[11]. The report argued that the advent of "killer robots", namely weapons systems that are capable of killing fully autonomously, was imminent. Since, according to "Losing Humanity", such systems would fail to comply with international humanitarian law, the report proposed a "preemptive prohibition on their use and development" (Docherty et al. 2012, p. 1). The report fed into a broadly anchored "Campaign to Stop Killer Robots" which was launched in April 2013.

At the same time, a report of the United Nations Special Rapporteur on extrajudicial, summary or arbitrary executions, Christoph Heyns, of April 2013 called for national moratoria on the development of "lethal autonomous robotics": "This report is a call for a pause, to allow serious and meaningful international engagement with this issue" (Heyns 2013, para. 33). This report had been preceded by an interim report by the previous UN Special Rapporteur on extrajudicial, summary or arbitrary execution, Philip Alston. Already the interim report diagnosed a lack of discussion in civil society about the employment of robots in warfare (Alston 2010, p. 16). The interim report had relied on Singer (2009), which had already broken some ground for a broader discussion about robots in warfare in general.

Heyns' report also recommended to the United Nations Human Rights Council to "call on all States to declare and implement national moratoria on at least the testing, production, assembly, transfer, acquisition, deployment and use of LARS [Lethal Autonomous Robotic Systems] until such time as an internationally agreed upon framework […] has been established" (Heyns 2013, para. 113). The report urged the establishment of an expert group to assess the implications of robots under humanitarian and human

[9] This chapter, especially subsection 3.1, draws on Burri (2016), where specific views of autonomy are examined.

[10] 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 (with Protocols I, II and III), 1342 UNTS 163 (engl.), 10 October 1980.

[11] For an immediate reaction to "Losing Humanity", see Thurnher (2013). For an accessible discussion of the process in Geneva, see Weaver (2014), pp. 131 *et seq*. For a broad discussion of robots and cybernetics, see Rid (2016).

rights law (ibid., para. 35). Soon thereafter, experts met three times informally in Geneva, namely in May 2014 (United Nations 2014), April 2015 (United Nations 2015), and April 2016 (United Nations 2016b) in the context of the CCW[12] to discuss lethal autonomous weapons and possible measures necessary to address them (on unmanned warfare more generally, see also Galliott 2016).

Autonomous weapons systems also came to figure prominently on the agendas of high-level events such as the World Economic Forum in Davos in 2016 (World Economic Forum 2016) and the Munich Security Conference of 2016 (Ignatius 2016). In December 2016, the fifth Review Conference of CCW formally decided to establish an open-ended Group of governmental experts to address lethal autonomous weapons:[13]

> *"To establish an open-ended Group of Governmental Experts (GGE) related to emerging technologies in the area of lethal autonomous weapons systems (LAWS) in the context of the objectives and purposes of the Convention, which shall meet for a period of ten days in 2017, adhering to the agreed recommendations contained in document CCW/CONF.V/2, and to submit a report to the 2017 Meeting of the High Contracting Parties to the Convention consistent with those recommendations. The GGE will hold its first session from 24 to 28 April 2017 or from 21 to 25 August 2017 and its second session from 13 to 17 November 2017 in Geneva. […]".*

A number of actors have converged around this CCW process, of which states have traditionally been in charge, in order to address autonomous weapons systems. Civil society is represented by NGOs such as Amnesty International (2015), Article 36, the International Committee for Robot Arms Control[14] and Human Rights Watch (Docherty et al. 2012, 2014, 2015). The international movement of the Red Cross, notably the International Committee of the Red Cross (ICRC 2016) and the International Conference of the Red Cross and the Red Crescent (ICRC 2015: pp. 44-47), is also effectively engaged in addressing autonomous weapons systems in the light of the laws of armed conflict.

The argument about autonomous weapons systems has also reached administrators, lawyers, and lawmakers concerned with the law applicable during peacetime (see the Dallas police incident discussed above, in section 2.2.2). Doubts have notably arisen whether the use of autonomous weapons systems by law enforcement would be in compliance with human rights, but the discussion is still in its infancy (Docherty et al. 2014; Heyns 2013). On a different note, organizations normally occupied with technical issues have become engaged in the discussion about law and ethics of autonomous systems more generally. Thus, the IEEE in 2016 began to devote time and effort to autonomous systems (including weapons systems). In a broadly anchored participatory process involving more than 100 experts the IEEE Global Standard Initiative on Ethical Considerations in Artificial Intelligence and Autonomous Systems elaborated a charter on "Ethically Aligned Design". While this charter discusses autonomous systems generally, it also includes a chapter on weapons systems (IEEE 2016, pp. 68-79). The European Parliament has begun to consider autonomous systems, too. It adopted a report with certain recommendations on civil use of

[12] See Final Report (United Nations 2013), Meeting of the High Contracting Parties to CCW, CCW/MSP/2013/10, 16 December 2013, para. 32 for the establishment of the first informal meeting of experts on lethal autonomous weapons systems; for the second meeting: Final Report, Meeting of the High Contracting Parties to CCW, CCW/MSP/2014/9, 27 November 2014, para. 36.

[13] Final Document (United Nations 2016) of the Fifth Review Conference [Advance version], FIFTH REVIEW CONFERENCE OF THE HIGH CONTRACTING PARTIES TO THE [CCW], 23 December 2016, Decision 1, p. 9.

[14] See www.icrac.net, including the "Original Mission Statement" of 2009 and the "Berlin Statement" of 2010 (under „Statements").

robotics.[15] This document does not address autonomous weapons systems in depth, though, for the European Union (and thus the Parliament) lacks the power to address security issues, since according to the founding treaties of the Union most powers in the security domain remain with the member states.

3.2 The Substance of the International Debate

The international debate about lethal autonomous weapons systems centers on the international humanitarian law laid down in the Geneva conventions (including the three additional protocols). The principal idea of international humanitarian law is that in armed conflict – in contrast to non-conflicts in which human rights apply rather than humanitarian law – it is not *a priori* unlawful to kill or harm humans provided that certain legal principles, namely distinction, necessity and proportionality, are observed. Roughly speaking, one may only attack military targets, such as combatants; the military gain an attack offers must also be in reasonable relationship to the damage it causes.

One of the main concerns under humanitarian law is that autonomous weapons systems will not be used in compliance with the principles of distinction and proportionality or that they undermine the human responsibility to ensure compliance with those legal obligations.[16] The application of these principles is highly context-dependent; a lawful target may become an unlawful target within seconds, e.g. when a combatant signals that he or she surrenders. (For further illustrations, see the case studies in section 2.3.1.) The principles also require the careful balancing of the interests at stake, sometimes in situations in which information is scarce and the time to consider a decision is limited. So far, humans have exercised measured judgment in these situations. Although the future development of technology is uncertain at this point in time, the worry now is that the use of autonomous weapons systems will not be in accordance with these principles and will consequently violate humanitarian law; hence the push by some for banning them.

It was described above in section 3.1 how civil society and non-governmental organizations drive much of the push for a ban. However, there is pushback too. Anderson and Waxman (2013), for instance, argue against a ban of lethal autonomous weapon systems, instead opting for an incremental approach by gradually evolving existing codes of conduct. According to Schmitt and Thurnher, a ban would be "insupportable as a matter of law, policy, and operational good sense" (Schmitt & Thurnher 2013, p. 233). Kerr and Szilagyi, in contrast, draw attention to the fact that lethal autonomous weapon systems would have an impact on international humanitarian law. Such weapons systems would notably result in a changed understanding of what would be considered militarily necessary. According to them, international humanitarian law contributes to this change in the idea of necessity through its neutrality towards new technology.

Despite this pushback, most stakeholders seem to agree in principle that the decision to kill should not be transferred to machines. The debate is most heated, however, on the question *when* this would happen,

[15] Draft Report with recommendation to the Commission on Civil Law Rules on Robotics, EUROPEAN PARLIAMENT (COMMITTEE ON LEGAL AFFAIRS; RAPPORTEUR: MADY DELVAUX), 2015/2103(INL), 31 May 2016. 396 MEP voted in favor, 123 against and 85 abstained (see Cooper & Plucinska 2017).

[16] This is not the only problem though. Roff (2014), identifies problems with the targeting process when autonomous weapons systems are involved.

while the uncertainty about the future "autonomy" of systems further complicates the debate. "Meaningful human control" over the decision to kill appears to be required, but it is hard to determine what this means.[17] Variations of control already exist in that humans may be "in", "on" or "out of the loop", but these are rough approximations to a wide range of control options available. (see also section 4.2.2.).

The law of armed conflict traditionally applies in physical, real-world conflict where embodied autonomous weapons systems can be deployed. However, autonomous systems also operate in cyberspace where the legal situation is even less clear (Walter 2015, p. 685). With cyberspace being outside the scope of this report, it is only noted briefly that the law of armed conflict and the prohibition to use force, a basic principle upon which the United Nations Charter is built (see article 2(4)), seem applicable in cyberspace (Schmitt 2013). In cyberspace, *control* (over software, etc.) is also a thorny legal problem, especially so when cyber warfare leads to loss of life.

If autonomous weapons systems were deployed in *non*-conflicts, e.g. for purposes of law enforcement and police work (again see the Dallas incidence above section 2.2.2, though the only dimension of autonomy it employed was "control distance", see section 1.3), the applicable legal framework would change (Asaro 2016). Depending on each state's treaty obligations, human rights law (rather than humanitarian law) is applicable; they are typically supplemented by constitutional rights and freedoms. Killing a human is only lawful under rare and exceptional circumstances within this civil human rights framework, so *lethal* autonomous weapons systems will likely feature less prominently than in armed conflict. However, the use of police force more generally may possibly be automated in the future. In the human rights framework, the question of when restrictions of the rights to life and corporal integrity resulting from the use of police force are lawful is answered by means of a careful exercise of balancing which involves flexible and highly context-dependent notions such as necessity and proportionality. The worry with autonomous systems is that they will not be capable of exercising the measured judgment needed for this balancing exercise in concrete situations. In addition, control over autonomous systems may be difficult to conceptualize and measure. While these issues broadly reflect the conceptual challenges autonomous weapons systems give rise to under international humanitarian law, the legal concerns under human rights law seem more serious, since collateral damage is tolerable to a lesser extent and strong due process rights need to be respected. It will be up to domestic legal orders to deal with these issues. In Switzerland, the cantonal police forces operate in a well-established and robust framework of cantonal administrative laws. It would have to be assessed with reference to the applicable cantonal law, for instance, whether the use of explosives in circumstances such as those prevailing in the Dallas incidence would be lawful. The fundamental rights individuals enjoy pursuant to the Swiss federal constitution (and international human rights instruments) would have to be fed into the assessment, namely the right to life and corporeal integrity (article 10 paragraphs 1 and 2 of the Swiss federal constitution). Precedent could provide some, though limited guidance.[18] To the best of the authors' knowledge, no case involving autonomous systems has so far reached the Swiss courts.

Under both humanitarian and human rights law, liability is a major concern.[19] With control distance increasing, causation, a key concept in liability, becomes difficult to establish. The behavior of autonomous systems may not always be fully predictable either, especially with machine learning systems (see section

[17] For attempts to conceptualize human control, see Crootof (2016, p. 9 et sq.), Roff & Moyes (2016), Sharkey (2016).

[18] See the famous killing shot taken by the police of the Grisons in Chur in 2000, discussed in Chapman 2010.

[19] Asaro (2011) already pointed out two problems with liability of robots/AI: moral agency and punishment. See also Human Rights Watch (Docherty et al. 2015) and von Bothmer, Frederik (2014).

1.1.8). Their behavior may sometimes not even be fully explainable even with the benefit of hindsight and under ideal circumstances of full transparency (see section 1.3 and Burri 2017). Consequently, it becomes unclear who is responsible when a system causes damage (on the responsibility gap, see below, section 4.1). While the resulting uncertainty complicates liability both under civil and criminal law, insurance is not a way out under criminal law (unless certain basic conceptions of criminal law were fundamentally changed). In international law, similar complications arise in the contexts of international criminal liability of individuals and international responsibility of states and international organizations. Certain technical approaches, though, seem to have the potential to make liability work for autonomous systems (Kroll et al. 2017, p. 699 et sq.).

A final legal problem with autonomous systems stems from the fact that they can be used in various ways. Certain uses may be relevant under humanitarian law, while others may be purely civilian. One typical example is a flying drone, which, depending on circumstances, may be used for civilian purposes or as a weapon. This potential dual use complicates legal assessment and enforcement (and a ban on autonomous weapons systems) as well as the testing of new weapons pursuant to article 36 of the Additional Protocol to the Geneva Conventions (Protocol I)[20]. For more details on testing and use of autonomous weapons systems, see US Department of Defense (2012).

3.3 Possible Developments in the Law

The outcome of the CCW process is uncertain at this point in time. Yet it is not very likely that a strong ban on autonomous weapons systems will be agreed upon. Operational autonomy along various dimensions (see section 1.3) and in various degrees is present in too many weapons systems already in lawful use for that to be a realistic option. "Autonomy" is also a less clear notion than, for instance, "blinding lasers". (A ban on blinding lasers had been the result of a process similar to that currently underway with regard to autonomous weapons systems; Doswald-Beck 1993). A ban would also have to be widely accepted in order to be effective. A legal instrument no one subscribes to or complies with is not in the interest of anyone involved in the CCW process.

A more likely outcome is the prohibition against or regulation of specific uses of autonomous weapons systems. Ideally, the notion of "meaningful human control" would be fleshed out in some more detail. This is by no means simple. The number of rules and regulations needed to allocate ordinary "control" at the right place in conventional armed forces (where no autonomous systems are involved) is indicative of difficulties ahead. Similarly, despite the evolution of a theory of agency in philosophy, political science and economics, we only have a basic understanding of who (or what) controls whom (or what) and under which circumstances.

The work of the group of governmental experts established pursuant to the decision taken in December 2016 in the context of CCW will take time. It took the international experts assembled by the ICRC in 1989 two years to come to grips with blinding lasers (and blinding lasers were relatively straightforward), and it took four more years for Protocol IV to the CCW on blinding laser weapons to be adopted. There is no guarantee that the work will end in success either. States will likely make their influence felt in the group

[20] Article 36 Protocol I: "New weapons. 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."

of experts and their interests diverge. The likely "haves" among them will be reluctant to forgo the strategic and tactical advantage autonomous weapons confer, while the likely "not-haves" will push for a strong ban. That the process will be wrenched from CCW – as happened with land mines when a separate, more expeditious negotiation channel had been opened which ultimately resulted in a comprehensive treaty banning land mines (Convention Mines 1997; for details see Burri 2016, pp. 357-358) – is a possibility. But there seems to be less consensus on autonomous weapons systems than there had been on land mines. Given the state of technology, it is probably also less urgent to regulate them. Finally, both ban and regulation could to some extent become redundant, if compliance with humanitarian law could somehow be programmed into autonomous systems (Arkin 2009; Anderson & Anderson Leigh 2015, p. 324 et sq.).

4 Ethics

In this chapter, we will review the ethics of autonomous robotic systems as defined in chapter 1 and in particular the definition of autonomy in section 1.3. The primary ethical concerns discussed in this section will relate to using autonomous robotics systems that are equipped with the capacity to harm or destroy. In the previous sections (sub-sections 2.3.1, 2.5.2, and chapter 3), we have discussed laws and policies as they relate to these systems but ethics is a third important factor in judging the value of autonomous weapons systems. Autonomous weapons systems are an emerging technology and emerging technologies add another layer of difficulty in our ethical analysis. The rate of change in these technologies can outpace the ability of governments to properly discuss and regulate them before the next iteration of change occurs (see the essays in Marchant et al. 2011). The disruption to the policy and legal status quo created by new technologies produces policy gaps where these new technologies can push what is possible beyond what was imaginable just a few decades earlier. The philosopher James Moor (2005) claims that these moments are where the philosophical study of ethics is particularly helpful. While ethics is neither a physical nor a social science, it is informed by law and the sciences and is a necessary first step in creating not only a survivable future with autonomous weapons systems, but also one that might be worth living in as well. We will begin by outlining the special ethical concerns raised in the discussions about autonomous robotics systems in general and autonomous weapons systems specifically. The next subsections will look at the ethical arguments that have been raised both pro and contra towards weapons systems autonomy. We will conclude with a short look at some likely developments in these technologies that must inform our discussion on the ethics of autonomous weapons systems.

4.1 Outlining the Ethics of Autonomous Robotics Discussion

Autonomous weapon systems raise many ethical concerns (cf. Enemark 2014; Galliott 2016; Di Nucci & Santoni de Sio 2016), as do autonomous vehicles aka self-driving cars (cf. Maurer et al. 2015) and healthcare/assistive/companionship robots (cf. Sharkey 2012a; van Wynsberghe 2015). Beyond the obvious political impacts there is also a growing apprehension regarding the likely effects that interacting with these kinds of systems will have on human social interaction (Turkle 2011; van den Brule et al. 2014). Others have focused on the risks of machine learning (Matthias 2004), especially when it is combined with systems autonomy (Haselager 2005; Noorman & Johnson 2014).

In order to begin to address these, and similar, concerns a close collaboration between science, engineering, and humanities has been called for (Veruggio & Operto 2008). This call has been heeded and the interdisciplinary study variously known as, Robot Ethics, Machine Ethics, or RoboEthics has come into being (Anderson & Anderson 2011; Lin et al. 2012; Royakkers & van Est 2015, Royakkers & van Est 2016). In this field, scientists and scholars from various disciplines actively work together towards the responsible development of robotics (Asaro 2008; Lichoki et al. 2011; Pagallo 2017). This work has grown out of the general methodology of embedding ethical and societal considerations into the development of technologies using frameworks like Responsible Innovation (Owen et al. 2012) and Value-Sensitive Design (Friedman & Kahn 2003; van den Hoven et al. 2007).

The ethical debate over autonomous systems can be divided in to the following main areas: responsibilities, rights, virtues, and harm. We will now look at each of these debates in turn. We remark that many of those debates are relying on experiences with the use of remotely piloted aircrafts (RPAs), which are not systems with a high degree of autonomy (see Chapa 2018). However, issues raised by RPA foreshadow the issues that might be raised by autonomous weapons systems.

4.1.1 The Responsibility Gap

A first set of concerns is whether autonomous systems may be designed in such a way that responsibility gaps between morally and legally relevant agents are avoided (Matthias 2004; Sparrow 2007; Human Right Watch, 2015). These relevant agents may be producers, owners, instructors, collaborators, or end-users for civilian applications; and producers, politicians, commanders or end-users for military applications. As robots cannot be expected to possess any high-level moral capacities, proper human accountability for a robot's actions has to be ensured at all times in morally, legally and economically appropriate ways (see the debate on military robots on this point in section 3.1 and Asaro 2008; Krishnan 2009; Guarini & Bello 2012; Sharkey 2007, 2012a, 2012b). The more autonomous capacities the systems possess, the more relevant this point becomes. This is a concern that cuts across many domains, but it has been more explicitly discussed in relation to military technologies and autonomous vehicles.

As was discussed in section 1.3, the sense in which weapons systems are said to be "autonomous" typically refers to the engineering sense of the word. However, we have to be careful since the term "autonomy" also has a long history in legal, moral, and ethical philosophy and equivocation in in the use of the word "autonomy" with between these various meanings is an easy and common mistake. Roff (2013) argues that maintaining an insistence on paying attention to the strong philosophical sense of autonomy is required in the moral analysis of lethal autonomous weapons systems, or we risk becoming completely befuddled when discussing what commentators refer to as the responsibility gap. Given the difference between autonomous robots and the autonomy required of moral agents, Roff prefers to use the term Lethal Autonomous Robot (LAR rather than the more common Lethal Autonomous Weapons (LAWs) or Lethal Autonomous Weapons Systems (LAWS). This is an important distinction as it makes LARs not simply one more type of weapons system, but instead they function more as a new kind of artificial combatant, one that has no means of making ethical decisions, and one that is not morally equivalent to the other human combatants it may engage with. These artificial agents can confuse our ability to ascribe responsibility for a given lethal act since the machine itself cannot be the locus of responsibility because in a robot, "there is no discussion of intent, of consciousness, or of the type of freedom that admits of moral operators of praise or blame. The robotic notion of autonomy is radically minimalist, as it removes ethical evaluation by definitional *fiat*" (Roff 2013).

Following this logic, LARs are not moral agents and therefore cannot be held to *jus in bello* considerations like any human soldier would, meaning that the more we see these systems used, the less we can hold anyone responsible for affronts that may occur to *jus in bello* that are committed by LARs:

> *The deployment of LARs in combat presents us with a never before seen challenge to just war theory. First, it divorces jus in bello judgments of responsibility from the behavior of combatants, as the combatants are no longer considered moral agents capable of moral standing. By doing so, it forces any evaluations of responsibility to jus ad bellum considerations of who decides to initiate war and to use LARs in combat. Instead of deciding whether political officials started an aggressive war, and thus can be charged with a crime of aggression, we must now discern whether those officials can be held morally and legally responsible for the conduct of autonomous machines (Roff 2013).*

This is a strong and important argument but it relies on a definition of moral agency that requires a robust requirement of freedom of will in determining which agents count as moral agents and which do not. It is possible that the requirement of robust freedom of will is too strong for even human agents given various biological, social, and cultural factors that certainly influence actions. If human agents cannot cross that very high bar, why should we expect robotic agents to do so as well before we grant them any moral status (Sullins 2006)? In this way if we place too high a requirement for free will in autonomous actions and accept Roff's critique of LARs, then we may be committing ourselves to a much deeper indictment of Just War Theory more generally since neither robot nor a human soldier can meet these requirements. Therefore, it follows that for pragmatic reasons we should consider the ramifications of these arguments but realize that none of them are yet conclusive.

We can see from this section that ascription of responsibility for the actions of Lethal Autonomous Systems is a difficult and open question, and this fact needs to be honestly discussed in any proceedings regarding the acquisition or use of any lethal autonomous system.

4.1.2 Human Rights and Autonomous Robotics Systems

A second concern is how to design systems so that they do not negatively affect basic human rights. This concern has been voiced especially in relation to possible violations of both physical and psychological integrity by healthcare robots (Sharkey and Sharkey 2012), violations of moral autonomy of medical patients (Sorell & Draper 2014), their privacy (Calo 2011) and dignity (Veruggio & Operto 2008; Coeckelbergh 2010; Sharkey 2014). In addition, autonomous military robots have been strongly criticized for their potentially negative impact on human rights as we discussed in section 3.1. Issues of rights and justice have also been raised in the debate on the crash avoidance algorithms to be embedded in self-driving cars (Lin 2015). Issues of privacy and security, such as the risks of these systems being hacked have been raised about surveillance drones. Furthermore, there are urgent concerns about the impact autonomous systems will have on employment and labor conditions and the right to work (e.g. Ford 2015).

4.1.3 Autonomous Robotics Systems and Human Virtues

A third set of concerns is whether and how autonomous robots may negatively affect the development of valuable human character traits, virtues, and skills. This phenomenon occurs in a similar way to the loss of professional skills, such as piloting skills, that become less frequently used in the presence of technologies such as automatic flight controls (for a good discussion of this see Mindel 2015). This concern includes the skills needed to establish meaningful human relationships both in specific professional contexts, such as care practices in medicine (e.g., van Wynsberghe 2015), and in ones that are more informal such as in friendship and loving relationships (e.g., Turkle 2015; Vallor 2016; Sullins 2012). The ethical debates on the impact of autonomous robots on human virtues, skills, and relationships impacts the discipline of social robots and affective computing but they can also have an effect on the design of military robots as well. Semi-autonomous robots have been criticized for potentially impeding the development of the traditional martial virtues and changing the perception of military valor – both the perception of military valor at a cultural level and the self-perception of the war-fighter themselves. Nick Turse and Tom Engelhardt for example wrote that utilizing these technologies "cannot be war, as anyone has ever understood the word, if one side is never in danger" (Turse and Engelhardt 2012, p. 64). And Noam Chomsky and Andre Vltchek wrote that "warfare has moved away from man-to-man combat, and is now dominated by deadly missiles, bombing campaigns and the latest terrible weapons: drones, which are synonymous with terrorism and absolute impunity – they kill without the invading nation having to risk its own soldiers" (Chomsky and Vltchek 2013, p. 172). In addition, militaries often pride themselves in the development of courage, bravery and valor under fire for their members. Thus, the question emerges whether war-fighters that are not physically present on the battlefield can share these virtues.

Given these concerns, it is likely that military virtues – that are highly prized by military organizations – will suffer through the expanded use of autonomous weapons systems. Courage, loyalty, honor, and mercy will all have to be substantially redefined in this new context (Sparrow 2013). Perhaps this will be seen as an acceptable loss, if it turns out that warfare using autonomous systems is less destructive and produces fewer casualties on all sides. We also have to acknowledge the possibility that these technologies might enable moral vice in the military context through fostering conflicts that are motivated by "colonialist, imperialist, or downright racist motivations…where the self-proclaimed *Herrenmenschen* and the harbingers of civilization discipline the brutes, mostly by killing them" (Steinhoff 2013, p. 207).

However, the position that military virtue will suffer under the effects of these new technologies does not go unchallenged. These worries have been challenged with the argument that these new technologies do provide suitable platforms for the development of martial virtue (Chapa 2014). Nevertheless, there is some evidence that, at least in the US military, there has been resistance from the military culture to giving awards and promotions to those service members that pilot remotely piloted aircrafts (RPA). However, there are those that argue that these criticisms do not speak to the reality of serving in this branch of the military and the very real stress it places on these warfighters (Blair 2012). Colonel Eric Mathewson[21] is reported to have said that "valor to me is not risking your life; valor is doing what is right. Valor is about your motivations and the ends that you seek. It is doing what is right for the right reasons. That to me is valor."

4.1.4 Moral Harm Caused by Autonomous Weapons Systems

The fourth set of moral concerns is the moral harm that is endured both by the warfighters who must do the killing that is demanded in wartime and those noncombatants who must live and survive in these new battle zones. We will address the moral harm to warfighters first. If it were possible to remove the threat of physical harm from warfighters through autonomous and semi-autonomous weapons systems, then it might be seen as an overall moral good. It is a reasonable and ethical choice to remove the soldier from the conflict. This move saves her from the physical risk and since they are not in direct combat, we might expect that there would be less of the psychological damage that can be done when we ask people to kill and then return them to civil society afterwards (Vallor 2016, p. 215). Nevertheless, it is not as simple as that. Researchers in the US Defense Department have reported that RPA aircrew[22] face problems such as depression, anxiety, and PTSD at the same rate as pilots that are flying manned aircraft in combat experience (Dao 2013; for more thorough psychological studies see Chappelle et al. 2012, 2013, 2014, 2014b; Christen et al. 2014; Fitzsimmons & Sangha 2013). Therefore, even if the aircrews of these machines are not in danger of physical harm, they are still very much in danger of suffering psychological and moral harm.

[21] Mathewson was Commander of the 432nd and 57th Operations Groups, Creech Air Force Base and Nellis Air Force Base, Nevada, where he led six active duty and six Air Reserve Component attack and reconnaissance squadrons flying the MQ-1 "Predator" and MQ-9 "Reaper" aircraft both in training and in combat operations around the globe. Information gathered from Colonel Mathewson Bio, retrieved from https://www.hpc.msstate.edu/UAS/files/Mathewson_Bio.pdf. The quote emerges from Jaffe, 2010, Combat Generation: Drone operators climb on the winds of change in the Air Force. Washington Post. February 28. http://www.washingtonpost.com/wp-dyn/content/article/2010/02/27/AR2010022703754.html

[22] Additionally, it is possible that some support personnel such as intelligence analysist involved in the operation of UAV may also suffer adverse psychological effects but at this time, this claim is disputed and lacks research to confirm or deny.

Obviously, there are also important effects to those societies that are under threat from these systems that we must also address. While these weapons do not produce the civilian casualties that traditional bombs might, they do produce a significant number of unintended casualties. Many of the persons targeted by these weapons tend to be irregular combatants that live amongst civilians. These civilians may or may not know that they are living next to a person whom is at war with states that can, and will, use lethal force against him or her thus placing their neighbors in mortal danger. Inevitably, people are caught in the blast radius who are not the intended target. When this happens, it can begin a tragic chain of events.

Since there has been a significant number of years of experience with semiautonomous weapons being used extensively in various conflicts, there has been some evidence of what it is like to live for years under constant surveillance which is punctuated by periodic deadly strikes. The International Human Rights and Conflict Resolution Clinic at the Stanford Law School with the Global Justice Clinic at the NYU Law School produced a report on the subject in 2012. The report suggests that there are some significant effects, "… mental health professionals, and journalists interviewed for this report described how the constant presence of US drones overhead leads to substantial levels of fear and stress in the civilian communities below" (International Human Rights and Conflict Resolution Clinic and Global Justice Clinic 2012, pp. 80-81). Since it is impossible to know who these drones may be tracking and targeting, civilians of all ages but children in particular who live in the areas of operation are reported to suffer from anticipatory anxiety which is common for those living in war zones(ibid). Those who may have seen the results of a drone strike often suffer from PTSD. We should remember that much of the time, these people are not living in official warzones, yet they suffer as if they were. This suffering includes physical manifestations of mental stress. One Pakistani physician interviewed for the report stated that he saw people suffering from, "…physical symptoms without a real [organic] basis, such as aches, and pains, vomiting, etcetera" (ibid).

A number of documentaries have attempted to track the effects of drone warfare on both warfighters as well as those living under the operations of these technologies. Two notable examples are "Living Beneath Drones," a documentary by Jamie Doran and Najibullah Quraishi (Doran and Quraishi 2016) and "National Bird," directed by Sonia Kennebeck (2016). Both of these films give the viewer a visceral look at how these technologies have effected individual lives.[23]

One tragic outcome of all this is when a relative of a casualty becomes motivated to take up arms as a result of civilian casualties even though they might not otherwise have done so. Often they are driven by the cultural need to seek revenge for the killing of a loved one. The West is not the only warrior culture involved in these conflicts and those in the blast radius have their own set of warrior virtues under which they try to live. Akbar Ahmed, in his book *The Thistle and the Drone*, warns us that when we inadvertently kill an innocent tribe member in the mountains of Pakistan while targeting a known terrorist, their kin have a moral debt to seek revenge: "A tribesman without honor is much like a Christian without a soul. Honor thus has social as much as spiritual content. Transgressions against honor necessitate revenge, which can often get out of hand. This brings to mind a Pukhtu proverb: 'He is not a Pukhtoon who does not give a blow in return for a pinch'" (Ahmed 2013, p. 53). However, since the perpetrator of killings that involve covert operations using RPA systems is not known and not even present at the time of the killing, there is no legitimate person for against whom the tribesman can seek revenge.[24] This can cause them to take targets of opportunity or to join forces with people that even they find morally reprehensible. Ahmed

[23] The fidelity of these documentaries have faced challenges. For an alternative analysis of "National Bird," see Chapa 2017.

[24] It is true that conventional aerial bombardment has many of these same problems, but with drone warfare, the strikes can happen well outside of known warzones and can often seem completely unprovoked to the civilian victims caught up in the killing.

(2013) claims that these tribesmen of Waziristan – after being confronted with drones – turned to a new type of weapon: the inhuman, un-Islamic, and deadly suicide bomber. It is claimed that this tactic of using suicide as an extension of war was unknown even during the bleakest days of fighting Soviet troops in Afghanistan (Ahmed 2013).

Thus, there is the possibility of serious moral harm that can occur to both the operators and the unintended victims of autonomous weapons systems. This negates the claim that autonomous weapons systems remove humans form the harmful effects of war. These problems are imminent in the case of asymmetrical warfare, where Shannon Vallor argues, "in fact, asymmetrical warfare arguably *doubles* the types of moral horror for which the human family is accountable. On the one hand, we have mechanical and impersonal technics of killing, one that fosters a military psychology in which the most grievous human suffering of innocent civilians can be reduced to the vocabulary and calculus of 'bugsplats'[25]. On the other side, we witness surges of increasingly gruesome and indiscriminate violence from those who pursue their low-tech methods of brutality without even the merest semblance of moral restraint" (Vallor 2016, p. 216). Those concerns need to be addressed in the design and deployment of future systems.

4.2 Ethics of System Autonomy

From the discussions so far in this report as well as in the sections that follow, we see that the primary issues that will determine the ethical and moral status of systems autonomy are as follows: Do the systems under analysis have meaningful human control? Are these systems safe and their use transparent? Do their learning algorithms lead to predictable behavior? Arguments against these systems all attempt to show that these systems display serious deficiencies in addressing these questions while the pro arguments will try to reassure us that each of these problem areas can be properly mitigated.

4.2.1 The Moral Status of System Autonomy

System autonomy is not a simple designation. As discussed in sections 1.3, a robot is a system composed of many subsystems, some of these subsystems might operate autonomously and others may not. The degree of autonomy a system displays is contextual and is a complex mix of human independence, mission complexity, and environmental complexity. Therefore, a system such as an RPA might be autonomous in certain contexts but not others. For instance, a RPA might operate under control of a remote aircrew but have the ability to return autonomously to a landing point if the system loses contact with its aircrew due to a malfunction or some kind of active jamming. Therefore, context and use will matter a great deal in the ethical analysis of autonomous systems. Likewise, a traditionally piloted aircraft might also be autonomous in some contexts. The US Air Force's C-17 Globemaster III, for example, is capable of flying entire missions by means of a pre-programmed flight management system (FMS).[26]

[25] The notion of 'bugsplats' appears widely in the critical literature about drones as an exemplar on the dehumanizing effect of this weapon. However, "bugsplat" was the name of a software that depicted the expected blast and fragmentation pattern of the various air-to-surface weapons in the U.S. inventory; it was renamed "Fast Assessment Strike Tool-Collateral Damage (FAST-CD)" already in 2003. It wasn't that "the dead" were depicted as squished bugs. People were not depicted at all. The software was developed to show how urban terrain would impact the blast and fragmentation pattern of a given weapon on a given target. See for further information: https://warontherocks.com/2017/06/drone-ethics-and-the-civil-military-gap/

[26] In fact, C-17 flight manuals encourage aircrew to fly using "the highest level of automation" (Air Force Instruction Manual 11-2C-17 2011, p. 73) available, so long as it does not act to the detriment of the crew's situational awareness.

Perhaps the most important aspect of the ethical analysis of autonomous systems comes in the capacity for that system to make its own plans of action using its own computational abilities. As we described in section 1.3., certain systems do not have the capacity yet others do. However, besides operational autonomy, we also need to address the fact that no system today explicitly includes ethical values in its decision-making process, no matter how artificial intelligent the system is. Systems that have an ethical impact but do not reason about that effect are sometimes called Ethical Impact Agents (EIA). While any system that reasons at all in any way about that impact could be called artificial ethical agents (AEA), and finally, a system that had the capacity to fully reason about the ethics and morality of its actions would be called an artificial moral agent (AMA). Wallach and Allen describe three types of artificial ethical agency in their book *Moral Machines* (2009). They break it down as a function of system autonomy and ethical sensitivity. For instance, a particular system might have a certain amount of autonomy in its decision making along with a fair amount of ethical sensitivity making it a viable AEA, or it might have a high amount of autonomy but a low level of ethical sensitivity making it a EIA and a dangerous one at that. If one considers artificial ethical agency as desirable, then the ultimate goal would be to build a system with high levels of both autonomous intelligence and ethical sensitivity making this system a full AMA. This ultimate goal is, however, beyond the technological capacities of today and those that are likely in the near future.

Following the suggestions of one of our interview partners (INTV06), these additional categories might be relevant for assessing the moral qualities of autonomous systems:

1) **Hybrid human 'control.'** We ought to have not only the physical manifestation of exoskeletal and other forms of human-machine hybrid limbs and joints; but furthermore we should have shared goal-seeking behavior by human-machine systems to relieve cognitive load on humans for a number of reasons[27].
2) **More robust social vision.** In-the-wild social vision is the ability for computational systems in any form to understand our social context, from emotion and tactical sociality to strategic social interactions around them. Autonomous systems would need such robust social vision. This will make computational systems far more socially 'aware' of the human group world around them.
3) **Autonomous help-request processes.** We will eventually need help centers available to provide human guidance to autonomous machines while the machine's primary operator is absent. For example, in areas like autonomous driving, when the driver is utterly gone, the system *must* be able to access a help center with humans in it that can provide help to cars. This pattern will be widespread, and for this to work, machines will need to have the general capability of understanding when to ask for help and how to ask for help correctly for situation awareness to happen properly.

This means that each of these potential autonomous systems has a different set of ethical and legal arguments for and against allowing autonomy within its field of operation and we will have to deal with this kind of complexity in our analysis. We will find that strong arguments can be made in favor of some systems autonomy where the ethical impact is low while others with high levels of ethical impact but low levels of ethical sensitivity are more deeply problematic.

[27] Hybrid human control is useful to relieve workload from humans, assist those who suffer from disabilities such as dementia, but we should also be wary of how they can contribute to the control of human behavior.

4.2.2 Meaningful Human Control

Human control of autonomous systems has been discussed at length in this report so far (1.1.4, 1.3, 3.2, 3.3). The notion of 'meaningful human control over individual attacks' was coined by the NGO Article 36, to express the core element that is challenged by the movement towards greater autonomy in weapons systems (Roff & Moyes 2016). This term was critiqued in section 3.2 as one that is still debatable in the legal context (see also Footnote 18). Strong proponents of meaningful human control such as the NGO Article 36 argue that it essentially means that sufficient human control is maintained in all aspects of the choosing of targets and the decision to use lethal force against those targets (Roff & Moyes 2016; Human Rights Watch (April 11,2016)). Two premises must be accepted to follow this argument, one is that it is unacceptable for a machine to apply lethal force without being under human control, and the second is that a human simply pressing a fire, or 'on' button is not sufficient to establish human control (Roff & Moyes 2016, p. 2). "On this basis, some human control is required and it must be in some way substantial – we use the term 'meaningful' to express that threshold" [28] (ibid). The Center for New American Security (Horowitz and Scharre 2015, p. 4) provides three criteria for meaningful human control:

1. Human operators are making informed, conscious decisions about the use of weapons.
2. Human operators have sufficient information to ensure the lawfulness of the action they are taking, given what they know about the target, the weapon, and the context for action.
3. The weapon is designed and tested, and human operators are properly trained, to ensure effective control over the use of the weapon.

From a legal standpoint, it follows that, if IHL explicitly requires human control of lethal weapons, then as soon as the degree of autonomy makes meaningful human control impossible autonomous weapons systems would not be in compliance. One might counter by stating that a requirement for human control is not explicitly stated anywhere in IHL but that it is merely implied by context or precedent. We have looked at the legal arguments in section 3 already. What still needs to be addressed is the claim that even if it can be decided that lethal autonomous weapons systems operating outside of meaningful human control are legal under IHL, we might still want to claim that they are nonetheless unethical or immoral in ways that the law does not yet capture. If one follows this argument, then lethal autonomous weapons systems highlight a potential ethical flaw in IHL, namely its inability to address the responsibility gap (see the discussion in section 4.1.1), and that deficiency needs to be addressed in the interests of ethical jurisprudence and to make the world more human and just. Some examples of those who make or discuss ethical arguments like these are; Sparrow (2007), Asaro (2012), Sharkey (2012b), Allen and Wallach (2013), Grut (2013), Roff (2013), O'Connell (2014), and Horowitz and Scharre (March 2015). One counter argument to this position is that there are no additional ethical problems that go beyond what IHL already addresses (Schmitt and Thurnher 2013), those who champion this position maintain that this debate is a legal issue alone to which ethics does not contribute. Another counter argument, which we will see again in the sections below, claims that meaningful human control might only mean that one is justly confident of the intended operation designed into lethal autonomous systems and that they will bring about just and legal outcomes when used (Arkin 2007, 2009, 2010; Lin et al. 2008; Anderson & Waxman 2013; Sullins 2010).

[28] It is important to note that not all the authors of this report endorse this argument, but it is an important one to know since it is plays a strong role in the case against autonomous weapons systems.

The above provides evidence that there is an intense debate on the moral status of systems autonomy and that debate will surface again in the sections below.

4.2.3 Major Ethical Positions in the Current Debate against Lethal Autonomous Weapons Systems

Researchers writing on the intersection of ethics and robotics have largely come out in opposition to the deployment of lethal autonomous weapons systems. Critics argue that as these systems become increasingly autonomous, the human designers and users of these systems commit an unjustified abdication of moral accountability in life and death decision-making (Altmann 2009; Asaro 2008; Sharkey 2009, 2010, 2011, and 2012; and Sparrow 2007, 2009a, 2009b, 2011). This has caused some of them to join an effort called *Stop Killer Robots*, which is a multinational effort to ban autonomous weapons systems from the battlefield (see section 3.1).

The other side of the debate rejects the *a priori* assumption that there is no accountability possible in these systems. This move treats accountability as an empirical problem, where we have to look at actual systems and analyze both the negative, as well as the potential positive ethical impacts they might have before we pass judgment on each system (Lin et al. 2008; Kershenar 2013; Schmitt 2013; Strawser 2010; Di Nucci & Santoni de Sio 2016). Still others argue that it might be possible to program higher levels of ethical sensitivity and value judgement in these systems (Arkin 2007, 2010; Sullins 2010). If either of these positions is correct, it is possible that not designing and deploying such systems could be itself an ethically problematic course of action given that these systems might have a more positive ethical impact on the battlefield than human warfighters would without the aid of these systems.

4.3 Arguments Contra System Autonomy

4.3.1 Autonomous Systems in General

Autonomous systems, even those that are not designed with the intention of causing harm, present certain concerns for those who argue against the use pf autonomous systems in general. The primary concerns we will look at in this section are; function creep, intransparency, implicit bias, responsibility, dual use, control, reliability, predictability, trust, and safety.

Function creep: As systems upgrade, surplus military equipment becomes available for use by civil authorities and this leads to the system being used in domains they may not have been initially designed for. Systems that are ethically tolerable in a military context are not always appropriate for civil society, and vice versa (Wynsberghe & Nagenborg 2016). For instance, systems designed to surveil dangerous terrorists that are later deployed in civilian contexts or in boarder control operations run grave risks of reducing civil privacy rights and might possibly run counter to human rights protections; e.g., when a country uses them to militarize a border where persistent illegal immigration might be occurring.

Intransparency and implicit bias: Many autonomous systems (such as RPAs, although they only have a low degree of autonomy) are first designed for surveillance purposes; i.e., they are not meant to draw attention to themselves. As the example of drones shows, it would be difficult for those living in a society that was using this technology to distinguish the normal commercial use of drones from those on surveillance missions. Additionally, those who might be aware that they are under surveillance are unlikely to be able to know for sure who is operating these systems and that would make it impossible for them to correctly judge if the actions these systems are doing are just and authorized by a legitimate authority. These systems may also dehumanize those being surveilled making the users of the surveillance drones less sensitive to honoring their human rights[29]. For instance, they would likely be used over populations that already have difficult relationships with law enforcement and this could be exacerbated by the implicit biases already present in the law enforcement agencies in question. One of our experts (INTV03) points out the problem that hidden/implicit harmful biases in training data/design/use might lead to unlawful, discriminatory or otherwise unjust outcomes for surveillance missions. A future example might be an earthquake disaster robot that prioritizes the rescue of victims from large homes or wealthy neighborhoods, or a security robot that disproportionately follows and harasses people of color at public gatherings. Thus, one would have to ensure that any self-learning system's program alterations are well understood and monitored for the emergence of unpredicted or novel harms.

Responsibility: Another problem is accounting for the responsible parties when these systems are involved in accidents. This can be seen already when hobby drones have made dangerous nuisances of themselves around fires or other emergency response situations, where they are trying to get photos of the disaster. It is difficult to determine who the operator is and this would be even harder if the system were operating with more autonomy. According to one of our experts (INTV05), intransparency of responsibility may not be an urgent requirement for non-weaponized systems. However, accidents could still happen, and there will need to be an accounting of that. Another complication for responsibility is the possibility that the system embeds the bias, either implicit or explicit of their designers. This could lead to an unethical diffusion of responsibility (INTV05).

Dual Use: Dual use occurs when the system can be used in many other ways than it was originally designed for, which may produce unintended consequences in its use and deployment. Systems designed to be nonlethal will not stay that way since they can be easily modified into lethal systems as was discussed in section 2. Generally, every system is lethal, even a self-driving supply vehicle, especially in a battle space. Machinery in general is potentially lethal (INTV04). Capacities that are the basic building blocks of autonomous systems prime them for use as lethal systems. As one of our expert said (INTV05): "It's a short hop from a nonlethal system to a lethal one, once the targeting/identification capability is in place."

One of our experts (INTV03) identified an interesting possibility that could be found if two, presumably ethically designed, systems were linked to make a third system whose ethical impacts had not been assessed: The linking of any seemingly autonomous but unarmed system with one that has complementary kinetic capacities that can take destructive action could be a tremendous temptation. For example, a human remotely piloted armed drone is used and we regard its principled and legal use as ethical. An AI-enabled facial recognition system installed in public spaces and trained to identify and track the position of likely hostile actors, with no kinetic powers, is another. However, it would be all too simple to link them in a network in which the human judgment and control of the lethal instrument is rendered practically meaningless, producing a functional equivalent of an unethical LAWS. We probably can best deal with these outcomes by trying to anticipate them in the design phase of these technologies; perhaps by using a value-sensitive design methodology that factors in potential ethical impacts and tries to mitigate them from the early stages of design (see INTV19 and Owen et al. 2012; Friedman et al 2003; van den Hoven et al. 2014).

[29] Privacy and surveillance are well-known ethical problems but autonomous systems present a novel vector for vast amounts of new data to be collected without the consent of those being surveilled.

Control, reliability, predictability, trust, and safety: Even non-weaponized systems that are made autonomous will have increased safety and control concerns. The more autonomous systems become, the larger these problems get. We might be required to tolerate some of these issues in certain circumstances but for most situations, we have to demand high levels of safety and reliability. All autonomous systems have to be predictable and it must be ethical to place our trust in them. When they will do the task they were designed for, we have to demand a high degree of success.

Autonomy may be very desirable in certain military applications where we want or have to limit contact with the system so that it is not discovered or hacked by an enemy, but that does not absolve us of the responsibility to be able to shut that machine down in the event that it is behaving in a way that was unforeseen. Autonomous technologies will need to be thoroughly audited and tested before it can be ethically deployed.

Certainly, all security sector technology must be audited/tested and its actual social effects studied and regulated to ensure that their design and use:

- Is safe
- Is secure (from abuse, hacking, etc.)
- Is more effective/less costly than traditional means
- Is reliable
- Allows appropriate role for human oversight and, if necessary, intervention
- Does not create, amplify or reinforce injustices, unlawful or immoral discrimination, or otherwise unjustifiably reduce human welfare (INTV03).

Furthermore, we have to realize that robotics is still a new technological field. While there have been significant advances in the success rates of robotic systems performing as programmed, we are still very far from the point at which these autonomous systems will be sufficiently safe to deploy. Videos that show off the great advances in robotics, particularly those videos released by Boston Dynamics, which display their ingenious walking robots that are built designed to support troops in the field in future conflicts[30]. However, one of our experts (INTV04) cautions us to remember that much of this is marketing. Initially Boston Dynamics systems failed all the time, they now fail half the time. This is a great improvement, but these systems would not be ethical to deploy at this time. When actual people's lives depend on it, these systems would have to work all the time.

A significant additional problem occurs when we contemplate adding machine learning capabilities to autonomous systems. Any system that used machine learning either in the design or programming phase, or that had the ability to learn from experiences in the field, would likely fail in an audit of its ethical use. This is because even the designers of a learning system have a difficult time fully describing why a system trained in this way behaves the way it does (for a good discussion on this see Burrell 2016). Because of this, it is difficult to ethically justify how an autonomous system might make a decision that has ethical implications. One of our experts (INTV05) remarks that in addition this problem as we do not understand how neural nets or certain learning algorithms in general work, not only are predictability and trust major issues, but liability becomes problematic as well.

[30] See the Boston Dynamics YouTube channel for many examples of these videos: https://www.youtube.com/user/BostonDynamics

4.3.2 Autonomous Weapons Systems

Autonomous weapons systems that are designed to be used in combat and law enforcement contexts cause harm to individuals and destroy property. This means that the ethical issues raised by them will have many different ethical impacts than autonomous systems in general. Yet some of the concerns discussed in the last section also apply here such as function creep, implicit bias, intransparency, accountability, control and safety. Additional concerns that we will address in this section include the ethical justification of allowing autonomous weapons systems to create physical harm to persons or their property whether that harm is intended to be lethal or nonlethal. Primarily these concerns center on whether or not we can believe that these systems are under meaningful human control, and on their capacity to comply with the Law of Armed Conflict (LOAC).

Function creep: This concern is shared with autonomous systems in general. However, it requires special mention here since it is a much more urgent concern with autonomous weapons systems than it is with systems designed for surveillance. While gradually creeping into a surveillance state is bad enough, one that regularly used armed drones with high degrees of autonomy in civil law enforcement would be completely intolerable. To the extent that this is a possibility, then this becomes a strong argument against designing and deploying these weapons systems.

Justifying physical harm: There are ethical issues in the use of systems that are designed to cause harm, whether the system is designed to cause lethal or non-lethal harm. In fact, some of the ethical concerns are exacerbated in the law enforcement context because in this context, human rights law and ethics dictate that the taking of life is typically unlawful and this is not entirely the same in the armed conflict context (see section 3.2). Thus, in either context, autonomous systems designed to cause harm generate special ethical concerns.

Above we discussed the legal challenges made by the Campaign to Stop Killer Robots (section 3.1) but here we will look more closely at the main ethical arguments raised against autonomous weapons systems. As ethical arguments, their force will depend upon their ability to convince you that a world with autonomous weapons systems is not one that is worth living in, even if it might be one in which these systems might be considered legal and/or beneficial for the survival of some culture or political system.

A final point in this argument is that at the very least machines should not be choosing targets in a military or security context. According to one expert (INTV19), unsupervised autonomous systems are unethical, since they lack the full capacities of human judgment, including for instance emotion, phronesis, and wisdom. This argument should cause us to maintain that choosing targets, even with non-lethal ammunitions, should still be something that is directed by a human controller. Another of our experts described the problem as one of accurate targeting (INTV05). According to him, systems have to be programed to correctly distinguish legitimate targets. However, in a world of guerilla warfare, where combatants do not wear uniforms, this might be impossible. If armed forces are tagged with identifying features such as RFIDs, for instance, they can be removed, or innocent civilians could be made to wear them to throw off targeting systems. Facial recognition systems also can be spoofed. Even identifying weapons may result in false positives, for example a shepherd with an AK-47 to protect his flock. The ethical problem here is that both humans and autonomous systems cannot meet the discrimination demands of just war theory and IHL.

Risk transfer and lowering the threshold for the use of lethal force: Earlier we discussed these concerns in detail in our discussion of autonomous systems in general. However, these problems take on a more tragic character when decision makers realize the political expediency of risk transfer. Since lethal danger is transferred from the human warfighters to machines, this can cause increased, and less thoughtful, deployment of lethal autonomous systems. When this happens there is far less political risk to those who decide to take us to war, thus potentially leading to more armed conflict (Strawser 2013). This problem follows into the civil use of these systems where the use of them may be chosen over more traditional policing methods that might be more expensive or difficult. This would lead to an increased use of lethal force over the capture of those suspected of crime.

Ethical justification/accountability: This is simply the acknowledgement that lethal decisions must be ethically justifiable. It is a serious decision to use force. Especially the use of intentional lethal force and there are important legal and ethical steps that must be taken before it is used. The worry here is that if we were to give autonomous systems the ability not only to make factual decisions but also value decisions, then, based on this argument, we have created a system that cannot be ethically justified. In order for the decision to be justified, we would have to be able to fully account for the actions of the machine not only from a mechanical description of how it chose a target but also why it was ethically justified in doing so (Wallach 2013).

Another of our experts (INTV19) elaborated further on this issue and it is worth looking at his comments in more detail. He begins with the claim that these systems are neither full artificial moral agents (AMA) nor do they deserve much consideration as moral patients, (meaning that they have no intrinsic rights that have to be respected beyond perhaps the property rights of their owners).[31] Given this ambiguous moral status, they should never be given autonomy for making decisions and taking actions that have a high level of ethical impact. Lethal capacity is only justified if a system has full capacities of human judgment and agency (and in addition full capacities of human patiency[32]), such systems should only be used under strict human supervision, and should not be fully autonomous, in particular with respect to the dimensions of control-distance and capacity for self-organization and collective agency. His claim is that humans rely on a certain set of skills that are peculiar to us and difficult, if not impossible, to replicate in machine intelligence. If capacities such as emotion, practical ethical wisdom, or the skill to notice and attend to ethical problems as they arise are indeed beyond machines, then we cannot say that the ethical decisions they make are equivalent to those made by a skilled human ethical agent.[33] According to the expert (INTV19), humans rely on emotions, practical ethical wisdom, and so on, whereas machines by definition lack these capacities, even if they could be very intelligent. Since only humans can experience the threats, risk, and suffering that comes with lethal threats, then machines cannot know what it means to (threaten to) kill or to (threaten to) harm a human being. Therefore, only human beings, if anyone or anything, should be allowed to make lethal decisions or commit lethal actions.

It is important to point out that these criticisms apply mostly to the autonomous targeting and killing of human targets. But a system that is designed to make ethical "calculations" might still be allowed to make

[31] For instance, they do not hold nor deserve to hold any rights similar to human rights this lack of a moral right to existence is one of the things that make them preferable to humans for putting into dangerous situations.

[32] The notion of "patiency" means the quality that humans have to be seen as moral patients and refers to their capacity to hold human rights, etc.

[33] One may have to distinguish two aspects here: either the delegation of moral decisions to machines is morally risky as they do not understand important facts and can make wrong decisions, or delegation is intrinsically morally wrong as life and death decisions should be made by agents equipped with ethical capacities, emotions, etc.

decisions not to fire or to make the decision to abort an action that fails to meet certain programed ethical parameters. This would produce systems that had a kind of functional morality that might be complex enough to make them an AEA or artificial ethical agent (see section 4.2.1). However, one can object that if a military action is proportional, discriminate, necessary and being taken in support of a just war, and the military action is intended to prevent some great harm (e.g., an enemy force firing upon civilians, for example), failing to act in these instances may itself be an ethical wrong. It follows that if this is the case, then it is not clear why one would allow the machine to commit an ethical wrong of one type (a decision not to fire causes moral harm), but of another type (a decision to fire causes moral harm).

Also, the expert (INTV19) does not exclude the possibility that such systems could have their own built-in ethical constraints, which could make "ethical" calculations that try to distinguish between civilian and non-civilian targets, etc. – i.e. they can and should have capabilities for functional morality. However, since these technologies are likely never sufficient in complex situations (and on the battlefield situations are generally complex) and since ethics should never be reduced to just following rules, calculation, or even what machine learning can do at this time, then humans should at least supervise and control all autonomous system (INTV19).

One could conclude from these arguments that a ban on lethal autonomous weapons is warranted and if a ban is not possible, then strong regulations should be enacted to control the deployment of these systems. This is actually what the expert recommends: "An international framework to regulate these weapons is absolutely necessary, morally speaking, and although a ban is not likely to happen, one should at least try to convince people/nations and influence their decisions" (INTV19).

Finally, two additional issues should be mentioned. The first is to acknowledge that there may be increased risk for non-combatants when autonomous weapons systems are deployed to the battlefield and if this were to happen, then it would be a strong reason against their deployment. The second is that some of the systems already in development are very complex and secret; if one of them were lost on a mission, there could be a strong motivation to get it back and this could lead to escalation of a conflict.

4.4 Arguments Pro System Autonomy

As chapter 1 has outlined, there are many good technological and operational reasons for improving autonomous capacities of systems; in particular, in settings, where the communication between operator and system may be unreliable and where humans may make more mistakes due to specific operating conditions (e.g., time constraints). Those technological reasons may also have ethical relevance; however, in the following, we focus solely on (positive) ethical considerations of the use of autonomous systems in the security sector.

4.4.1 Autonomous Systems in General

As we will see in this section, some of the factors that worry opponents of autonomous weapons systems such as function creep, dual use, and risk transfer can take on a positive role in the arguments that favor the development and use of autonomous systems in general.

Function creep and dual use: Not all function creep is necessarily a bad thing. A good example is the use of surveillance systems in the area of conservation and environmental monitoring. Autonomous systems could do a lot of good in protecting endangered species and the environment. One expert (INTV06) mentioned that 'giving chase' to poachers is something that might work extremely well with such systems. E.g., non-lethal chasing of rhino hunters, persistently tracking them, or anyone else that is a suspect will probably become ubiquitous across countries with threatened animal populations, and these systems can be highly automated.

Earlier we brought up the concern that these systems might contribute to making borders more militant. However, they can also help in decreasing deaths that occur as people attempt to illegally cross the border in dangerous terrain. One expert (INTV06) mentioned the example that far fewer people might die of thirst after getting lost crossing the Texas border, because the systems may detect them and order help to rescue them.

These environmental and humanitarian missions might not be what the systems were originally designed to do but both military and civilian entities can extend the missions of these systems to include these tasks.

Inevitability: It is clear that across the globe autonomy is going to continue to be a hot research topic. As discussed in section 1, autonomy is far too useful, in so many different realms, not to be researched and deployed. This means that countries that wish to participate in these technological developments have to find a way to make the best of these coming technologies and find ways of dealing with the problems that may emerge in the various application domains. One expert (INTV05) remarked that even high-profile accidents could be treated as outlier cases that need to be debugged, rather than cautionary lessons to reduce autonomy.

Risk Transfer: In the section on the arguments against autonomous systems above, we saw that risk transfer could be seen as ethically worrisome, but there are important factors that can make this a pro argument. Since these technologies allow war-fighters, police, and rescue personnel to be removed from potentially harmful situations. This is an undeniable appeal and means that we can expect a great deal of research dedicated to making this a reality since it helps keep warfighters from physical harm. As one expert clarified (INTV04), autonomy is a pillar of US Department of Defense (DoD) planning and we will be seeing a lot of it in the future. Many of the applications will be non-lethal. Anything that reduces danger to war-fighters is a priority.

4.4.2 Autonomous Weapons Systems

Many of the arguments in favor of autonomous systems in general also apply to autonomous weapons systems. There are, however, a few additional arguments that specifically apply to autonomous weapons.

Limit allowed targets: One way around the serious ethical problems we discussed above is to limit the targets we allow these systems to engage. Our experts have some suggestions on how this might work. One expert believes that combat situations involving robots vs. robots will emerge first, at least at scale, to demonstrate reliability and control in lethal actions (INTV05). Much of the public outrage around the systems already in use (RPAs and the like, although those systems only have very limited autonomy) might be because they are used in ways that produce civilian casualties. Acceptance of lethal autonomous systems might be enhanced by deploying them in less ethically challenging situations. There is very little public outcry against bomb disposal robots, for instance. As one of our experts observes they are best used not in situations around civilians such as counter terrorism or counter insurgency. Examples include patrol in demilitarized zones, perimeter protection, use in monitoring nuclear weapons and facilities in verifying arms reduction treaties, as well as being used in building clearing operations. These uses are not without ethical impacts but it is easier to make the case that they are ethically permissible (INTV04). The argument here is that autonomy of various degrees has been a part of a number of weapons systems and increasing system autonomy should be promoted in ways that will not invite undue criticism. For instance, undersea systems that loiter and wait for a target, or any other fire and forget systems. Autonomy in this sense has been around for years (INVT04). They act in environments where a low degree of autonomy already is considered ethically acceptable – and increasing system autonomy in those circumstances is less likely to create ethical problematic outcomes.

Control and Safety: One of the more important pro arguments is that as long as lethal autonomous systems result in making situations on the modern battlefield that are at least as ethical as they are now, then they are a useful technology (Lin et al 2008). Arkin (2007, 2009, 2010, 2014) makes the case that ethical behavior by human soldiers on the battlefield is not as good as we would hope for. This means that while lethal autonomous systems might not be ideal in regards to the ethical impacts they might cause, these outcomes are still likely to be preferable to those that human agents might cause in similar situations. In our interview with Arkin he elaborated on this point explaining that in regards to lethal actions from autonomous weapons systems their "[e]thical impact must be at a minimum as good as a human but they should actually be better than human ethical decisions" (INTV04). Some of our experts also added that, "It is important to explain what "meaningful human control" means. This should include the ability of the system to cause the human users to make better ethical decision than they would have without the system (INTV04).[34]

Ethical Accountability and responsibility: In the section on contra arguments above, some critics felt strongly that lethal autonomous weapons were an abdication of the human responsibility to make lethal decisions. Even though the role of the human might be reduced down to simply turning the machine on and pointing it at an enemy the loss of accountability and responsibility can be seen as a red herring fallacy. In truth the responsibility lays where it always has, with the commanders and political leaders that chose to deploy these systems. One of our experts expressed this idea thusly, "Whomever gave the system its target signature, the author of its orders, these people can be held responsible" (INTV04). Whether we can simply decide to make individual people responsible in this way and in this context, is a very controversial issue both from a moral and from a legal point of view (see, for instance, Saxon 2016). However, it is true that there are plenty of real human moral agents involved in the design and deployment of these systems so what the systems do is ultimately the result of human decision-making – so no abdication of accountability or responsibility for the actions of the systems would be legitimate. Given that these systems collect a vast amount of data during their operation, it is possible that these systems might actually increase accountability and responsibility by leaving a thick data trail that can be analyzed after every mission. "[Autonomy should provide] the ability to identify and understand nuances in many scenarios that could weigh against lethality, i.e., to avoid incorrect kill decisions. But if [the systems] are made overly cautious, then operators may lose trust in the systems, making them less useful" (INV05).

International Humanitarian Law (IHL): As we discussed in chapter 3 and section 4.2.1 in detail, we have a large body of settled law in the IHL that provides specific guidelines for the use of lethal force. According to Schmitt (2013), if we diligently follow these laws, then the use of lethal autonomous systems will

[34] See sections 3.2 and 4.2.1 for a discussion on "meaningful human control".

be legally justified and therefore ethically justified as well (please refer to section 4.2.1 for more discussion). The pro argument here suggests that it is better to regulate these weapons through IHL than try to ban them outright and lose any influence over those that would build them anyway. However, one of our expert suspects that it is unclear whether additional international laws, even a ban or a moratorium, can be enforced or verified (INTV05). It is one thing to inspect a nuclear or chemical facility, and another to inspect any company that's develops robotics or AI.

Another benefit provided by these systems is that human warfighters have a very natural tendency to prefer to preserve their own lives, even if that preference might lead to a regrettable outcome such as firing when they perceive a weapon that is not there. Lethal autonomous weapons systems would not need to have preference strong motive for self-preservation. They could wait to apply lethal force and fully assess the situation before acting, even if that risks their destruction. This would allow them to have an initial stance against using force, which is something that is hard to achieve with human warfighters or even human police officers. One of our experts argues that this gives the autonomous system an important ethical advantage over humans. "Even in humans, the choice not to shoot is the cornerstone of their own autonomy and ethical behavior (INTV04).

At this time, the world is not engaged in any conflicts between superpowers, but if human history is any indication of the future, it is just a matter of time before it could happen again. In the same way that the international community was unprepared for the atrocities committed in WWI and WWII, we might be in the same situation regarding the technologies that could be used in another large-scale conflict. "This is why we have to solve these problems now, and not wait until political situations force us to deal with weapons we are unprepared to deal with" (INTV04). Following this line of thought, Arkin (2007, 2009, 2010) therefore argues against a ban on the research and development of lethal autonomous weapons: "A ban on these systems may not be the best thing for us since these systems are not entirely defined. Sure, let us ban 'the Terminator' but we do not have those, we have specific systems that need to be dealt with on a case-by-case basis. Some of these systems may indeed provide moral good that we will not receive if there is a blanket ban" (INTV04).

4.5 Likely Developments

The two major trends to come out of the discussions in this section are calls for: a) machine ethics (can we teach morality to machines?), and b) Value-sensitive design (how can we design systems to prevent unwanted consequences of autonomy while reaping its potential benefits?)

Developments in machine ethics will be necessary to address the growing autonomy already evident in "smart" weapons systems. In section 4.1.1, we saw that an important part of the debate is whether or not artificial agents can be considered moral agents. In section 4.2.1 we discussed the relationship between ethical awareness and autonomy described in Wallach and Allen (2009). The dangerous option is to create systems that have a high degree of autonomy but a low degree of ethical reasoning capacity and then place that system in a situation where it will significantly impact humans. On the other hand, if machine ethics is successful, then these systems will be autonomous ethical agents (AEA) which would have the capacity to better navigate these ethically significant situations.

Developments in values sensitive design (VSD) techniques are growing in the civilian context. A significant example of this is the EU Data Protection Regulation that has created "Data Protection by Design and by Default"[35]. We can see law and ethics working together here to create real change in the way technologies are designed and deployed. It is uncertain if VSD will be implemented in defense systems design but it should be encouraged through legal and ethical arguments and government policies.

Some additional technological developments may have an impact on the ethical assessment of system autonomy. We tend to think of these future autonomous systems in terms of what we already have, which is single systems like the MQ-1 Predator that may interact with other weapons systems. However, current RPAs only have a low degree of autonomous capacities comparable to any other 21st century military aircraft. They may therefore misguide our intuitions. For example, in the near future, we are likely to see autonomous systems operating in swarms. These swarms will likely have emergent properties that come about as multiple systems attempt to engage a target. Furthermore, as autonomy becomes more complex, it is actually going to be difficult to classify autonomous systems in a way that would facilitate regulations or bans by treaty. While biological weapons are easy to define, autonomy presents more dimensions for legal and ethical assessment and resists clear definitions.

One of our interviewees, the roboticist Ronald Arkin, described his recent work on advancing systems autonomy for this report. He and his research group have been working on autonomy for over thirty years in all manner of domains including military and civilian. One of their ongoing projects is the development of *Slow*-Bots, which are modeled after biological systems such as the Slow Loris. They are learning what they can about how this primate survives and they are applying that knowledge to create autonomous systems that can persist in the environment for years at a time. These kinds of systems could be used for many applications but one military application would be persistent surveillance. *Slow-bots*, would have very low energy requirements depending on the application they are used for, and would be able to reap that energy from the environment itself. They might also require very low levels of communication with human users and the communications they do have with human users will be accomplished through autonomously formed ad-hoc networks of robots (Arkin 2014). Often systems autonomy is conflated with the idea that the machines will have to be large and complex but as this example shows, it may first come with modest machines that have very low energy requirements and operate mostly through ad hoc networks that autonomously work together to accomplish the task they are intended to do.

There also is likely to be advancement on larger more traditional systems as well. As we have seen in the previous section, all of our experts place the requirement that these systems must have some capacity to reason ethically if they are to be responsibly deployed in a situation where they were to make lethal decisions. On the other hand, many of the experts are deeply skeptical that this can be achieved anytime in the near future. For example, the roboticist Illah Nourvakhsh does not believe that the whole "robots thinking ethically" issue will pan out in the next decade. People will realize that rules of engagement are nowhere near rational or codifiable enough given the actual perceptual abilities of machines, for machines to implement them. Instead, also in the near future we will see highly shared control with humans-in-the-loop for lethality (INTV06).

Unlike law, some ethical systems resist being put into a code. Ethical norms are not always simple and are constantly evolving as well as being deeply contextual. This means that there is not a book or code that one may point to that could be translated into programing and used by an autonomous system. The

[35] See: http://www.eudataprotectionregulation.com/data-protection-design-by-default [accessed October 24 2017].

roboticists we interviewed both cast doubts that we are anywhere close to being able to program ethics into a machine (INTV04, INTV06). However, behaving ethically and being ethical are very different things. Even in the far future, we may never be able to be an actual ethical agent with full moral reasoning capabilities, but the capacity for these systems to behave ethically is within reach. It is not unreasonable to expect that near future developments that will take the current status of our lethal autonomous weapons systems form ethical impact agent (EIA) to autonomous ethical agent (AEA), even if full Artificial Moral agency (AEA) is out of reach.

5 Material

5.1 Author Team

Thomas Burri is assistant professor of international law and European law at the University of St. Gallen in Switzerland; Dr. iur. (Zurich), LL.M. (College of Europe, Bruges), lic.iur. (Basel), admitted to the bar of the canton of Zurich, Venia Legendi for international law, European law, and constitutional law. Contact: Thomas.burri@unisg.ch

Joseph Chapa is a Major in the U.S. Air Force and an instructor in the U.S. Air Force Academy Department of Philosophy, Colorado/USA. Currently, he is doctoral student in philosophy at the University of Oxford/UK. His research interests are in Applied Ethics, the Just War Tradition and Remote Warfare. Contact: joseph.chapa@us.af.mil. *The views expressed are those of the author and do not necessarily reflect those of the US Air Force, the US Department of Defense, or the US Government.*

Markus Christen is research group leader at the Institute of Biomedical Ethics and History of Medicine and Managing Director of the UZH Digital Society Initiative of the University of Zurich, Switzerland. He is the principal investigator of this project. His research interests are in empirical ethics, neuroethics, and ethics of technology. Contact: christen@ethik.uzh.ch

Raphael Salvi is assistant at the Chair for Philosophy, ETH Zurich, Switzerland. His research interests lie in the fields of philosophy of technology and ethics of specific emerging technologies. Contact: raphael.salvi@phil.gess.ethz.ch

Filippo Santoni de Sio is assistant professor at the Section Ethics/Philosophy of Technology of Delft University of Technology, the Netherlands; he is co-director of the NWO research project "Meaningful Human Control over Automated Driving Systems". His research interests are in the theory of moral and legal responsibility, and in robot ethics. Contact: f.santonidesio@tudelft.nl

John Sullins is full professor in the department of Philosophy and a board member of the Sonoma State University Center for Ethics Law and Society, California/USA. His specializations are: philosophy of technology, philosophical issues of artificial intelligence/robotics, cognitive science, philosophy of science, engineering ethics, and computer ethics. Contact: john.sullins@sonoma.edu

5.2 List of Interviewed Experts

The following experts were interviewed (face-to-face, Skype or e-mail) for the preparation of this report. Not all expert interviews are directly referenced. All interviewed experts were invited to comment the draft of the report; experts marked with a (*) provided feedback.

INTV01 Nicholas Mull, US Navy, former head of operations in the law department (he was in charge of the article 36 weapons review for the US Navy). Interview on Saturday, 8 October 2016 by Thomas Burri.

INTV02 Julian Padget*, Senior Lecturer at the Dept. of Computer Science, University of Bath, UK (he is involved in the IEEE Global Standard Initiative on Ethical Considerations in Artificial Intelligence and Autonomous Systems). Interview on Wednesday, 21 December 2016 by Thomas Burri.

INTV03 Shannon Vallor*, Department of Philosophy, Santa Clara University, USA (she is President of the international Society for Philosophy and Technology). Interview on Wednesday, 28 December 2016 by John Sullins.

INTV04 Ronald Arkin, School of Interactive Computing, Georgia Tech, Director of the Mobile Robot Laboratory (he is a leading expert in the growing field of robot ethics). Interview on Monday, 14 November 2016 by John Sullins.

INTV05 Patrick Lin, Philosophy Department and Director, Ethics + Emerging Sciences Group, California Polytechnic State University, USA (he is well published in technology ethics, especially on robotics and AI). Interview on Saturday, 11 November 2016 by John Sullins.

INTV06 Illah Nourbakhsh, Department of Philosophy, The Robotics Institute at Carnegie Mellon University, USA (she is Director of the Community Robotics, Education and Technology Empowerment (CREATE) lab). Interview on Sunday, 1 January 2017 by John Sullins.

INTV07 Travis Burdine (Colonel, US Air Force retired), Strategy and Business Development Manager at Insitu, Inc., A Boeing Company (he is involved in the design & production of small and medium sized unmanned aircraft systems for defense, government, and commercial applications). Interview on Wednesday, 23 November 2016 by Joe Chapa.

INTV08 Heather Roff*, Senior Research Fellow in the Department of Politics and International Relations, University of Oxford and a Research Scientist at the Global Security Initiative, Arizona State University, USA (she is an Artificial Intelligence, Future of War, and Cyber Security Fellow at the New America Foundation). Interview on Thursday, 10 November 2016 by Joe Chapa.

INTV09 Matthew Studley*, Associate Head of Department, Engineering Design and Mathematics and Head of Research and Scholarship at the University of The West of England (UWE), Bristol, UK (he is a senior member of the Bristol Robotic Lab). Interview on Thursday, 25 November 2016 by Joe Chapa.

INTV10 Michael Meier, Special Assistant to the US Army Judge Advocate General for Law of War Matters (he is a retired US Army Officer and Judge Advocate and former attorney at the US Department of State). Interview on Monday, 21 November 2016 by Joe Chapa.

INTV11 Reto Wollenmann, deputy head arms control, disarmament, non-proliferation, EDA & Michael Siegrist, international law, EDA (both work for the Federal Department of Foreign Affairs (EDA) and are involved in the United Nations CCW discussion on Lethal Autonomous Weapons Systems). Interview on 6 December 2016 by Markus Christen & Raphael Salvi.

INTV12 Roland Siegwart, Director of Autonomous Systems Lab, ETH Zurich (he is a leading researcher in the field of autonomous robotics). Interview on 13 December 2016 by Markus Christen & Raphael Salvi.

INTV13 Daniel Krauer, Col., Swiss Armed Forces, Doctrine Research, VTG / Anita Noli-Kilchenmann*, Doctrine Research, Swiss Armed Forces, VTG / Martin Krummenacher, Doctrine Research, Swiss Armed Forces, VTG (Federal department of Defence, Civil Protection and Sports). Interview on 12 December 2016 by Markus Christen & Raphael Salvi.

INTV14 Marco Hutter, Robotic Systems Lab ETH Zurich (he is a researcher in autonomous robotics). Interview on 30 November 2016 by Markus Christen & Raphael Salvi.

INTV15 Neil Davison*, Scientific and Policy Adviser, Arms Unit, Legal Division, International Committee of the Red Cross, Arms Unit (he provides technical and policy advice on weapons issues in support of the ICRC's priorities to better protect victims of armed conflict). Interview on 5 January by Markus Christen.

INTV16 Vincent Choffat, Deputy Head Arms Control and Disarmament, Swiss Armed Forces (he is a specialist in Arms Control and disarmament and follow the discussion at the UN in Geneva). Interview on 21 December 2016 by Markus Christen & Raphael Salvi.

INTV17 Robert Babuska, Scientific director of the TU Delft Robotics Institute. Interview on 1 December 2016 by Filippo Santoni de Sio.

INTV18 Chris Verhoeven, Associate professor, department of microelectronics at the Delft University of Technology (he is Leader of the Swarm Theme of the TU-Delft Robotics Institute). Interview on 12 January 2017 by Filippo Santoni de Sio.

INTV19 Mark Coeckelbergh, Department of Philosophy, University of Vienna, Austria (he is member of the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems). Interview on Wednesday, 11 January 2017 by John Sullins.

INTV20 John Blitch, Department of Behavioral Sciences and Leadership, US Air Force Academy, Colorado (he is an artificial intelligence researcher with the Air Force Research Laboratory). Interview Friday, 16 December 2016 by Joe Chapa.

INTV21 Anonymous; Special Agent, United States Federal Bureau of Investigations (FBI). Interview by Joe Chapa.

5.3 List of Workshop Participants

The following experts were invited for a workshop together with the research team that took place January 19-20 2017 in Zurich. During the workshop, a first draft of the label system has been discussed. All invited experts received the draft of the report; experts marked with a (*) provided feedback.

- Anita Noli-Kilchenmann*, Swiss Armed Forces, Doctrine Research
- Daniel Messelken*, University of Zurich; Centre for Ethics, Center for Military Medical Ethics
- Heather Roff*, University of Oxford and Arizona State University, Global Security Initiative
- Markus Hoepflinger*, armasuisse Research+Technology, Research Director autonomous systems
- Martin Krummenacher, Swiss Armed Forces, Doctrine Research
- Matthew Studley*, University of the West of England, Dept. Engineering, Design and Mathematics
- Michael Siegrist, Federal Department of Foreign Affairs, Humanitarian Rights Section
- Neil Davison*, International Committee of the Red Cross, Arms Unit (day 1 of the workshop only)
- Ron Arkin, Georgia Institute of Technology, Director Mobile Robot Laboratory
- Vincent Choffat, Federal Department of Defense, Deputy Head Arms Control and Disarmament

5.4 List of Abbreviations

9/11	September 11 (2001) attacks
AEA	Artificial Ethical Agent
AI	Artificial Intelligence
AK-47	Awtomat Kalashnikowa aka Kalashnikov (Russian Assault Rifle)
AMA	Artificial Moral Agent
BCI	Brain-Computer Interface
CBRNC	Chemical, Biological, Radiological, Nuclear and Explosives
CCW	Certain Conventional Weapons Convention (UN)
CRASAR	Center for Robot-Assisted Search and Rescue
DMZ	Demilitarized Zone
DoD	(United States) Department of Defense
EDA	Federal Department of Foreign Affairs (CH)
EIA	Ethical Impact Agent
GGE	Group of Governmental Experts
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
HEL	High Energy Laser
HVI	High Value Individual
ICBM	Intercontinental Ballistic Missile
ICRC	International Committee of the Red Cross
IED	Improvised Explosive Device
IEEE	Institute of Electrical and Electronics Engineers
IHL	International Humanitarian Law
INTV	Interview
ISR	Intelligence, Surveillance, Reconnaissance
LAR	Lethal Autonomous Robotics
LAWS	Lethal Autonomous Weapons Systems
LEL	Low Energy Laser
Lidar	Light detection and ranging
LOAC	Law of Armed Conflict
MIT	Massachusetts Institute of Technology
NGO	Non-Governmental Organization
PED	Processing, Exploitation and Dissemination
PTSD	Posttraumatic Stress Disorder
RFID	Radio-Frequency Identification
RPA	Remotely Piloted Aircraft
UAV	Unmanned Aerial Vehicle
UGV	Unmanned Ground Vehicle
UMV	Unmanned Maritime Vehicle
UN	United Nations
US	United States of America
VSD	Value Sensitive Design
VTG	Federal Department of Defence, Civil Protection and Sports (CH)
WAMI	Wide Area Motion Imagery

5.5 Annotated Literature

Below, the cited literature is listed. Some important papers are accompanied with a short comment outlining their content (printed in ***bold italic***).

Air Force Instruction Manual 11-2C-17 (2011): *C-17 Operations Procedures*, Vol. 3, 16 November 2011 (incorporating Change 1, 20 March 2015) pp. 73. URL: http://static.e-publishing.af.mil/production/1/af_a3_5/publication/afi11-2c-17v3/afi11-2c-17v3.pdf [accessed October 24 2017].

Akbar, Ahmed S. (2013): *The Thistle and the Drone: How America's War on Terror Became a Global War on Tribal Islam*, Brookings Institution Press: Washington.

Allison, Brendan Z. et al. (2007): *Brain–computer interface systems: progress and prospects*, in: *Expert Review of Medical Devices*, Vol. 4, No. 4, pp. 463-474.

Alston, Philip (2010): *Interim report of the Special Rapporteur on extrajudicial, summary and arbitrary executions*, submitted to the Human Rights Council, focuses "especially on the relevance of new technologies in tackling the challenge of extrajudicial executions and the rampant impunity that attaches to the phenomenon", August 23 2010, United Nations (UN) General Assembly: Geneva. ***The interim report diagnosed a lack of discussion in civil society about the employment of robots in warfare: "Although robotic or unmanned weapons technology has developed at astonishing rates, the public debate over the legal, ethical and moral issues arising from its use is at a very early stage, and very little consideration has been given to the international legal framework necessary for dealing with the resulting issues." This UN interim report had relied on Singer, Peter W. (2009): Wired for War.***

Altmann, J. (2009): *Preventive Arms Control for Uninhabited Military Vehicles*, in: *Ethics and Robotics*, R. Capurro and M. Nagenborg (eds.), AKA Verlag: Heidelberg.

Amnesty International (2015): *Autonomous Weapons Systems: Five Key Human Rights Issues for Consideration*, Amnesty International Publications: London.

Anderson, Michael and Anderson, Susan Leigh (2011): *Machine Ethics*, Cambridge University Press: New York.

Anderson, Michael and Anderson, Susan Leigh (2015): *Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm*, in: *Industrial Robot: An International Journal*, Vol. 42, No. 4, pp. 324-331.

Anderson, Kenneth and Waxman, Matthew C. (2013): *Law and Ethics for Autonomous Weapon Systems: Why a Ban Won`t Work and How the Laws of War Can*, in: *American University, WCL Research Paper 2013-11* and *Columbia Public Law Research Paper 13-351*, Stanford University, The Hoover Institution (Jean Perkins Task Force on National Security and Law Essay Series***). These authors argue against a ban of lethal autonomous weapon systems, instead opting for an incremental approach by gradually evolving existing codes of conduct.***

Aqel, Mohammed O.A. et al. (2016): *Review of visual odometry: types, approaches, challenges, and applications*, in: *SpringerPlus*, 5:1897.

Arkin, Ronald C. (2007): *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*, Technical Report GIT-GVU-07-11, Mobile Robot Laboratory College of Computing GATECH: Atlanta.

Arkin, Ronald C. (2009): *Governing Lethal Behavior in Autonomous Robots*, CRC Press: Boca Raton.

Arkin, Ronald C. (2010): *The case for ethical autonomy in unmanned systems*, in: *Journal of Military Ethics*, 9(4): 332–341.

Arkin, Ronald C. (2014): *Bio-inspired Slowness for Robotic Systems*, Mobile Robot Laboratory College of Computing GATECH: Atlanta, URL: http://www.cc.gatech.edu/ai/robot-lab/online-publications/Arob2.pdf [accessed March 28 2017].

Article 36 Non-governmental organization (NGO): *The organization specifically also deals with the subject of autonomous (lethal) weapons and has spoken at the CCW informal meetings, published articles about meaningful human control in autonomous weapons systems and is considered as a founding member of the "Campaign to Stop Killer Robots". The website of the non-profit organization holds an extensive documentation. URL: http://www.article36.org [accessed March 27 2017].*

Asaro, Peter (2008): *How Just Could a Robot War Be?*, in: *Current Issues in Computing and Philosophy* (pp. 50-64), P. Brey, A. Briggle, & K. Waelbers (eds.), los Press: Amsterdam.

Asaro, Peter (2011): *A Body to Kick, But Still No Soul to Damn: Legal Perspectives on Robotics*, in: *Robot Ethics: The Ethical and Social Implications of Robotics*, pp. 169-186, Lin, Patrick et al. (eds.), MIT Press: Cambridge. ***This Paper points out two key problems with liability of robots/AI: moral agency and punishment.***

Asaro, Peter (2012): *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making*, in: *International Review of the Red Cross*, Special Issue on New Technologies and Warfare, Summer 2012, Vol. 94, No. 886, pp. 257-269.

Asaro, Peter (2016): *"Hands Up, Don't Shoot!" HRI and the Automation of Police Use of Force*, in: *Journal of Human-Robot Interaction*, Vol. 5, No. 3, pp. 55-69. This paper is focusing on human robot-interfaces for police work. ***The author urges a moratorium on the automated use of force in police work due to the serious challenges one is confronted with when it comes to "automating violence".***

Baldi, Pierre (2016), as cited in: Castelvecchi, Davide (2016): *The black box of AI*, in: *Nature*, Feature News, Vol 538, Oct. 2016, p. 222: Macmillan/Springer.

Beer, Randall D. (2009): in: *Scholarpedia* 4(4):1531, revision #91061.

Berman, Lazar (2012): *Israel's Iron Dome: Why America is Investing Hundreds of Millions of Dollars*, American Enterprise Institute (AEI), September 24 2012, URL: https://www.aei.org/publication/israels-iron-dome-why-america-is-investing-hundreds-of-millions-of-dollars/ [accessed March 21 2017].

Blair, Dave (2012): *Ten Thousand Feet and Ten Thousand Miles: Reconciling our Air Force culture to Remotely Piloted Aircraft and the New Nature of Aerial Combat*, in: *Air and Space Power Journal*, Vol. 26, No. 3, May-Jone 2012, pp. 61-69.

Bohidar, S. et al. (2014): *Energy Supply System in Robotic Machines*, in: *IJIRST*, Vol. 1, Issue 6.

Borton, David A. et al. (2013): *An implantable wireless neural interface for recording cortical circuit dynamics in moving primates*, in: *Journal of Neural Engineering*, April 2013, Vol. 10, No. 2: 026010.

Burgess, Lisa (2008) *Report faults computer in Guam B-2 crash*, in: *Stars and Stripes*, 7 June 2008, URL: https://www.stripes.com/news/report-faults-computer-in-guam-b-2-crash-1.79781#.WcJmUZOGOuU [accessed September 20 2017].

Burrell, Jenna (2016): *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, in: *Big Data & Society*, Januar-June 2016, Vol. 3, No. 1, pp. 1-12.

Burri, Thomas (2016): *The Politics of Robot Autonomy*, in: *European Journal of Risk Regulation*, Vol. 7, No. 2, June 2016, pp. 341-360.

Burri, Thomas (2017): *Machine Learning and the Law: 5 Theses*, URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2927625 [accessed March 24 2017].

Caliskan-Islam, A. et al. (2016): *Semantics derived automatically from language corpora necessarily contain human biases*, in: ArXiv.org, eprint arXiv:1608.07187.

Calo, Ryan M. (2011): *The Drone as Privacy Catalyst*, in: *Stanford Law Review*, December 2011, Vol. 64, No. 29, SLR Online, URL: https://www.stanfordlawreview.org/online/the-drone-as-privacy-catalyst/ [accessed April 22 2017].

Campaign to Stop Killer Robots: One of the prominent campaigns with its goal to "pre-emptively ban fully autonomous weapons". The campaign is an international coalition with some well-known members in their steering committee (such as Human Rights Watch, Article 36, ICRAC, Nobel Women's Initiative etc.), mostly NGOs. URL: http://www.stopkillerrobots.org [accessed March 28 2017].

Castelvecchi, Davide (2016): *The black box of AI*, in: *Nature*, Feature News, Vol 538, Oct. 2016, pp. 20-23: Macmillan/Springer. ***This publication gives a clearly understandable introduction on what problems one is confronted with in machine learning – especially when confronted with "deep learning" that is already implemented in a wide range of applications.***

Center for Robot-Assisted Search and Rescue (CRASAR): Texas A&M University, URL: http://crasar.org [accessed March 31 2017].

Chapa, Joseph (2014): *The Virtuous Drone Pilot*, MA Thesis, Boston College: Boston, URL: http://hdl.handle.net/2345/bc-ir:104048 [accessed April 22 2017].

Chapa, Joseph (2017): Film Review: *National Bird, Directed by Sonia Kennebeck*, in: *Journal of Military Ethics* Vol. 16, Nos. 1-2, pp. 130-137.

Chapa, Joseph (2018): *The Ethics of Remote Weapons: Reapers, Red Herrings, and a Real Problem*, in: *One Nation Under Drones*, ed. John Jackson (Naval Institute Press, forthcoming 2018).

Chapman, Matthias (2010): *Die Problematik bei einem finalen Rettungsschuss*, in: *Tages Anzeiger*, September 10 2010, URL: http://www.tagesanzeiger.ch/schweiz/standard/Die-Problematik-bei-einem-finalen-Rettungsschuss/story/30715061 [accessed March 22 2017].

Chappelle, Wayne L. et al. (2013): *Symptoms of Psychological Distress and Post-Traumatic Stress Disorder in United States Air Force "Drone" Operators*, in: *Military Medicine*, Vol. 179, No. 8, pp. 63-70.

Chappelle, Wayne L. et al. (2012): *Prevalence of High Emotional Distress and Symptoms of Post-Traumatic Stress Disorder in U.S. Air Force Active Duty Remotely Piloted Aircraft Operators*, 2010 USA-FAM Survey Results, Final Technical Report, Air Force Research Laboratory: Wright Patterson AFB.

Chappelle, Wayne L. et al. (2014): *Assessment of Occupational Burnout in United States Air Force Predator/Reaper 'Drone' Operators*, in: *Military Psychology*, Vol. 26, No. 5-6, pp. 376-385.

Chappelle, Wayne, Tanya Goodman, Laura Reardon, William Thompson (2014) *An Analysis of Post-Traumatic Stress Symptoms in United States Air Force Drone Operators*, in: *Journal of Anxiety Disorders*, June 2014, Vol. 28, No. 5, pp. 480-487.

Chomsky, Noam and Vltcheck, Andre (2013): *On Western Terrorism: From Hiroshima to Drone Warfare*, Pluto Press: London.

Christen, Markus et al. (2014): *Measuring the Moral Impact of Operating 'Drones' on Pilots in Com-bat, Disaster Management and Surveillance*, Twenty Second European Conference on Information Systems: Tel Aviv.

Coeckelbergh, Mark (2010): *Health care, capabilities, and AI assistive technologies*, in *Ethical Theory Moral Practice*, Vol. 13, No. 2, pp. 181–190.

Convention Mines (1997): *Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction*: 2056 UNTS 241; 36 ILM 1507 (1997).

Cooper, Harry and Plucinska, Joanna (2017): *Don't kill us, R2-D2: MEPs warn against robot revolt*, in: *Politico*, February 16 2017 resp. February 21 2017, URL: www.politico.eu/article/dont-kill-us-r2-d2-meps-warn-against-robot-revolt [accessed March 22 2017].

Cornwall, Warren (2015): *In Pursuit of the Perfect Power Suit*, in: *Science*, October 16 2015, Vol. 350, No. 6258.

Crootof, Rebecca (2016): *A Meaningful Floor for "Meaningful Human Control"*, in: *Temple International and Comparative Law Journal*, Vol 30, pp. 53-62.

Dao, James (2013): *Drone Pilots Are Found to Get Stress Disorders Much as Those in Combat Do*, in: *The New York Times*, February 22 2013, URL: http://www.nytimes.com/2013/02/23/us/drone-pilots-found-to-get-stress-disorders-much-as-those-in-combat-do.html [accessed March 27 2017].

DC Metropolitan Police Department, *Automated Speed Enforcement: Frequently Asked Questions*, URL: http://mpdc.dc.gov/page/automated-speed-enforcement-faq [accessed 11 Jan 2017].

De Crook, Arthur (ed.) et. al. (2016): *Robotics for Future Presidents: Leading Experts on the Next Revolution in Automation*, Robotics Institute, TU Delft: Delft.

Defense Science Board (2016*): Summer Study on Autonomy*. US Department of Defense, June 2016. URL: http://www.acq.osd.mil/dsb/reports/2010s/DSBSS15.pdf [accessed October 22 2017].

De Maria, G. et al. (2012): *Force/tactile sensor for robotic applications*, in: *Sensors and Actuators A: Physical*, 175. 60-72: Elsevier.

Del Prado, Guia Marie (2015): *This drone is one of the most secretive weapons in the world*, in: *Business Insider UK, Tech News*, September 29 2015, URL: http://uk.businessinsider.com/british-taranis-drone-first-autonomous-weapon-2015-9 [accessed March 21 2017].

Di Nucci, Ezio and Santoni de Sio, Filippo (2016): *Drones and responsibility: legal, philosophical and socio-technical perspectives*, Routledge: London.

Docherty, Bonnie et al. (2012): *Losing Humanity: The Case against Killer Robots*, Human Rights Watch and Harvard International Human Rights Clinic Report, Goose, Steve (ed.), November 2012: New York. ***With its main goal to protect human rights of people worldwide, HRW published one of the first NGO reports in 2012 that deals directly with the subject of "fully autonomous weapons" and their legal and ethical concerns. They acknowledge that fully autonomous systems do not yet exist, but argue that such systems would be unable to meet basic principles of international humanitarian law, and consequentially recommend (together with The International Human Rights Clinic (IHRC) at Harvard Law School) to all states that the use, development and production of fully autonomous weapons should pre-emptively be prohibited "through an international legally binding instrument".***

Docherty, Bonnie et al. (2014): *Shaking the Foundations: The Human Rights Implications of Killer Robots,* Human Rights Watch and Harvard International Human Rights Clinic Report, Goose, Steve (ed.), May 2014: New York.

Docherty, Bonnie et al. (2015): *Mind the Gap – the Lack of Accountability for Killer Robots,* Human Rights Watch and Harvard International Human Rights Clinic Report, Goose, Steve (ed.), April 2015: New York. ***This report goes more deeply into the subject of accountability issue when it comes to the use of fully autonomous weapons or "killer robots". It argues that, even if there is success in liability assignment, "the nature of the accountability that resulted might not realize the aims of deterring future harm and providing retributive justice to victims." The report supports the case against fully autonomous weapons and the call for a ban (see also: Docherty, Bonnie et al. (2012): "Loosing Humanity: The Case against Killer Robots").***

Dorigo, Marco et al. (2014): *Swarm robotics*, in: *Scholarpedia*, 9(1):1463, revision #138643.

Doran, Jamie and Quraishi, Najibullah (2016): *Living Beneath Drones*, Al Jazeera, September 19.

Doswald-Beck, Louise (1993): *Les armes qui aveuglent – rapports des réunions d'experts organisées par le Comité Internationale de la Croix Rouge sur les lasers de combat 1989-1991*, CICR: Geneva.

Drezner, Jeffrey A. and Leonard, Robert S. (2002): *Innovative Development: Global Hawk and DarkStar - HAE UAV ACTD Program Description and Comparative Analysis*, RAND Corporation: Santa Monica.

Enemark, Christian (2014**):** *Armed Drones and the Ethics of War: Military virtue in a post-heroic age*, Routledge: London.

European Parliament (2016): *Draft Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(NL))*, PE 582.443v01-00, Committee on Legal Affairs, Rapporteur: Mady Delvaux, May 31 2016.

Fetz, Eberhard E. (2015): *Restoring motor function with bidirectional neural interfaces*, in: *Progress in Brain Research*, Vol 218, pp. 241-252, Elsevier: Amsterdam.

Fitzsimmons, Scott and Karina Sangha (2013) *Killing in High Definition: Combat Stress among Operators of Remotely Piloted Aircraft*, Canadian Political Science Association/American Political Science Association; International Studies Association Annual Meeting.

Ford, Martin (2015): *Rise of the Robots: Technology and the Threat of a Jobless Future*, Basic Books: Boulder. See also: Kaplan, Jerry (2015).

Friedman, B., and Kahn, P. H., JR. (2003): *Human values, ethics, and design*, in: *The Human-Computer Interaction Handbook*, J. Jacko and A. Sears (Eds), Lawrence Erlbaum Associates: Mahwah NJ.

Galliott, Jai (2016): *Military Robots - Mapping the Moral Landscape*, Routledge: New York.

Gibbons-Neff, Thomas (2016): *ISIS used an armed drone to kill two Kurdish fighters and wound French troops, report says*, in: *The Washington Post*, October 11 2016, URL: https://www.washingtonpost.com/news/checkpoint/wp/2016/10/11/isis-used-an-armed-drone-to-kill-two-kurdish-fighters-and-wound-french-troops-report-says/?utm_term=.891c8d960b77 [accessed March 22 2017].

Grut, Chantal (2013): *The Challenge of Autonomous Lethal Weapon Systems to International Law*, in: *Journal of Conflict and Security Law*, Vol. 18, No. 1, pp. 5-23. ***This article discusses the implications of weapons autonomy on IHL.***

Guarini, Marcello and Bello, Paul (2012): *Robotic warfare: Some Challenges in Moving from Noncivilian to Civilian Theaters*, in: *Robot Ethics: The Ethical and Social Implications of Robotics*, pp. 129-145, Lin, Patrick et al. (eds.), MIT Press: Cambridge.

Guo, Yi (ed.) (2013): *Selected Topics in Micro/Nano-robotics for Biomedical Applications*, Springer: New York.

Haar, Rohini J. & Iacopino, Vincent (2016): *Lethal in Disguise: The Health Consequences of Crowd-Control Weapons,* Zwibel C., Suciu A., Ennarah, K., Pol, L., Santos, L. (eds.), Report of the International Network of Civil Liberties Organizations and Physicians for Human Rights.

Haselager, Willem F.G. (2005): *Robotics, philosophy and the problems of autonomy*, in: *Pragmatics & Cognition*, Vol. 13, No. 3, pp. 515-532.

Heyns, Christof (2013): *Annual report of the Special Rapporteur on extrajudicial, summary and arbitrary executions*, submitted to the Human Rights Council pursuant to its Resolution 17/5, focuses on lethal autonomous robotics and the protection of life, April 9 2013, United Nations (UN) General Assembly: Geneva. ***Christof Heyns, the then UN Special Rapporteur on extrajudicial, summary and arbitrary executions, submitted his annual report to the Human Rights Council of the United Nations that focuses on lethal autonomous robotic systems and the protection of human life. It recommends to states around the world a "national moratoria" on aspects of lethal autonomous robotic (LAR) systems, and "calls for the establishment of a high level panel on LARs to articulate a policy for the international community on the issue." While doing this, the report raises questions about how such systems can be programmed to comply with IHL and how legal accountability is dealt with.***

Hickman, William B. (MG) (2015): *Investigation Report of the Airstrike on the Médicins Sans Frontiéres / Doctors Without Borders Trauma Center in Kunduz, Afghanistan on 3 October 15*, declassified, USFOR Afghanistan: Kabul.

Horowitz, Michael C. and Scharre, Paul (2015): *Meaningful human control in weapon systems: A Primer*, Project on Ethical Autonomy, Working Paper, March 2015, Center for New American Security.

Human Rights Watch (April 11, 2016): *Killer Robots and the Concept of Meaningful Human Control: Memorandum to Convention on Conventional Weapons (CCW) Delegates*, HRW and IHRS, April 2015. ***A discussion of what has changed in regards to human control when we move from standard weapons to ones that have autonomous capabilities.***

Human Right Watch (2015). Mind the Gap: The Lack of Accountability for Killer Robots. Retrieved from https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots

Ienca, Marcello and Haselager, Pim (2016): *Hacking the brain: brain–computer interfacing technology and the ethics of neurosecurity*, in: *Ethics and Information Technology*, Vol. 18, No. 2, pp. 117-129.

Ignatius, David (2016): *In Munich, a frightening preview of the rise of killer robots*, in: *The Washington Post*, February 16 2016, URL: https://www.washingtonpost.com/news/checkpoint/wp/2016/10/11/isis-used-an-armed-drone-to-kill-two-kurdish-fighters-and-wound-french-troops-report-says/?utm_term=.891c8d960b77 [accessed March 22 2017].

Institute of Electrical and Electronics Engineer (IEEE) (2016): *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems*, Version 1 for Public Discussion, December 13 2016: United States.

International Committee for Robot Arms Control (ICRAC): Founded in 2009 by Juergen Altmann, Peter Asaro, Noel Sharkey and Rob Sparrow, this International Non Governmental Organization (NGO) "seeks to discuss with the international community implications of the use of robotics and strive for a "regulation of robot weapons" and a "peaceful use of robotics in the service of humanity". Mission statement(s) URL: http://icrac.net/statements [accessed March 31 2017].

International Committee of the Red Cross (ICRC) (2006): *A Guide to the Legal Review of New Weapons, Means, and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977*, ICRC: Geneva.

International Committee of the Red Cross (ICRC) (2015): *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, 32nd International Conference Paper, October 2015, Red Cross and Red Crescent Conference (December 8-10 2015), ICRC: Geneva

International Committee of the Red Cross (ICRC) (2016): *Autonomous Weapon Systems - Implications of Increasing Autonomy in the Critical Functions of Weapons*, ICRC Report, Expert Meeting, March 15-16 2016: Versoix.

International Committee of the Red Cross (ICRC Customary): *Customary IHL Rules*, in: *Customary IHL Database* URL: https://ihl-databases.icrc.org/customary-ihl/eng/docs/Home [accessed March 22 2017].

International Human Rights and Conflict Resolution Clinic and Global Justice Clinic (2012): Living under Drones: *Death, Injury, and Trauma to Civilians from US Drone Practices in Pakistan*, Stanford Law School & NYU School of Law: September 2012.

ISR - Deputy Chief of Staff for Intelligence, Surveillance, and reconnaissance (2016): *Small Unmanned Aircraft Systems (SUAS) Flight Plan: 2016-2036, Bridging the Gap Between Tactical and Strategic,* Report, U.S. Air Force, April 30 2016: Washington.

Jaffe, Greg (2010): *Combat Generation: Drone operators climb on winds of change in the Air Force*, in: *The Washington Post*, February 28 2010, URL: http://www.washingtonpost.com/wp-dyn/content/article/2010/02/27/AR2010022703754.html [accessed March 28 2017].

Judson, Jen (2016): *US Army Putting Finishing Touches on Autonomous Systems Strategy*, in: *DefenseNews*, March 17 2016, URL: http://www.defensenews.com/story/defense/show-daily/ausa-global-force/2016/03/17/army-autonomous-system-strategy/81897736/_[accessed March 22 2017].

Kaag, John and Kreps, Sarah (2014): *Drone Warfare*, Polity Press: Cambridge.

Kaplan, Jerry (2015): *Humans need not apply: A guide to wealth and work in the age of artificial intelligence*, Yale University Press: New Heaven.

Karimi, Faith, Shoichet, Catherine E., Ellis, Ralph (2016): *Dallas Sniper Attack: 5 Officers Killed, Suspect identified*, in: *CNN Online*, July 9 2016, URL: http://edition.cnn.com/2016/07/08/us/philando-castile-alton-sterling-protests/ [accessed March 21 2017].

Kennebeck, Sonia (2016): *National Bird*, Public Broadcasting Station (PBS), Independent Lens series, aired 1 May 2017.

Kerr, Ian and Szilagyi, Katie (2016): *Asleep at the switch? How killer robots become a force multiplier of military necessity*, in: *Robot Law*, Calo, R.; Froomkin, A; M., Kerr I. (eds.), pp. 333-366**. *For the authors autonomous military robots ("killer robots") are "force multipliers" and therefore have the potential not only for increasing "destructiveness and fatalities" when developed, but also have the ability to "change our own perceptions of "necessity and proportionality" and therefore have an impact in international humanitarian law.***

Kershenar, Stephen (2013): *Autonomous Weapons Pose No Moral Problems*, in: *Killing By Remote Control: The Ethics of an Unmanned Military*, Strawser, Bradley Jay, (eds.), Oxford University Press: Oxford.

Killmister, Suzy (2008): "Remote Weaponry: The Ethical Implications," *Journal of Applied Philosophy* Vol. 25, No. 2, pp. 121-133.

Kraska, Peter B. & Cubellis, Louis J. (1997): *Militarizing mayberry and beyond: Making sense of American paramilitary policing*, in: *Justice Quarterly*, Vol. 14, No. 4, pp. 607-629.

Krishnan, Armin (2009): Killer Robots: Legality and Ethicality of Autonomous Weapons. Ashgate Publishing, Ltd.

Kroll, Joshua A. et al. (2017): *Accountable Algorithms*, in: *University of Pennsylvania Law Review*, Vol. 165, pp. 633-705.

Liberatore, V. et al. (2004): *Robotic Communication Systems for Flexible, Sustainable, Affordable, and Autonomous Space Operations*, *A white paper prepared for NASA*, Case School of Engineering, Ohio.

Lichocki, P., Kahn, P., & Billard, A. (2011): *The ethical landscape of robotics*, in: *Robotics & Automation Magazine*, IEEE, Vol. 18 No. 1, pp. 39-50.

Lin, Patrick (2015): *Why Ethics Matters for Autonomous Cars*, in: *Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte*, Maurer M., Gerdes J. Ch., Lenz B., Winner H. (eds), Springer Open.

Lin, Patrick et al. (2008): *Autonomous Military Robotics: Risk, Ethics, and Design*, California Polytechnic State University: San Luis Obispo.

Lin, Patrick et al. (2008): *The Ethical and Social Implications of Robotics*, MIT Press: Cambridge.

Lin, Patrick et al. (2012): *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press: Cambridge.

Malenic, Marina (2016): *USAF 'Loyal Wingman' UAVs to emerge within a decade*, in: *IHS Jane's 360, May 19 2016, URL: http://www.janes.com/article/60471/usaf-loyal-wingman-uavs-to-emerge-within-a-decade* [accessed March 22 2017].

Marchant, G., et al. (2011): *The Growing Gap Between Emerging Technologies and Legal-Ethical Over-sight: The Pacing Problem*, Springer Netherlands: Dordrecht.

Matthias, Andreas (2004): *The responsibility gap: Ascribing responsibility for the actions of learning automata*, in: *Ethics and Information Technology*, September 2004, Vol. 6, No. 3, pp. 175-183.

Maurer, Markus, et al. (2015): *Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte*, Maurer M., Gerdes J. Ch., Lenz B., Winner H. (eds), Springer Open.

McNally, David (2014): *Autonomous System Development: The U.S. Army Prepares for The Future*, in: *Army Technology*, October 24 2014, Vol. 2, No. 6, p. 10.

Menthe, Lance et al. (2012): *The Future of Air Force Motion Imagery Exploitation: Lessons from The Commercial World,* Technical Report, RAND Corporation: Santa Monica.

Mindel, David A. (2015): *Our robots, Ourselves: Robotics and the Myths of Autonomy*, Viking: New York.

Mitchell, Tom M. (1997): *Machine Learning*, Series in Computer Science, McGraw-Hill: New York. ***The author gives an explanatory introduction to the machine learning approaches that includes examples and discussion of relevant algorithms, techniques and concepts. It tackles some basic concepts that are important in the fields of "artificial intelligence" and can be seen as one of the early standard work on machine learning.***

Moor, James (2005): *Why We Need Better Ethics for Emerging Technologies*, in: *Ethics and Information Technology*, Vol. 7, No. 3, pp. 111–119. Reprinted in: van den Hoven (2008), pp. 26–39.

Murphy, Robin B. (2004): *Activities of the Rescue Robots at the World Trade Center from 11–21 September 2001*, in: *IEEE Robotics & Automation Magazine*, September 2004, Vol. 11, No. 3, pp. 50-61.

National Aeronautics and Space Administration (NASA) (2015): *NASA Technology Roadmaps - TA 4: Robotics and Autonomous Systems.* ***While the NASA's Technology Roadmap on Robotics and Autonomous System is generally a public available decision support tool for NASA to***

estimate technology candidates for their future investments and applications, it also can give an outlook on what technologies in these fields most likely will develop within the next approx. twenty years and what obstacles such developments may encounter from a Space Agencies view that is more and more confronted with Robotics and Autonomous Systems.

Norman, Merel and Johnson, Deborah G. (2014): *Negotiating autonomy and responsibility in military robots*, in: *Ethics and Information Technology*, March 2014, Vol. 16, No. 1, pp. 51-62.

O'Connell, Mary Ellen (2014): Banning Autonomous Killing: *The Legal and Ethical Requirement That Humans Make Near-Time Lethal Decisions*, in: *The American Way of Bombing*: *Changing Ethical and Legal Norms from Flying Fortresses to Drones*, Evangelista M. & Shue H. (eds.), Cornell University Press: Ithaca/London. A discussion of the lawful and ethical problems associated with lethal autonomous weapons.

Owen, Richard et al. (2012): *Responsible research and innovation: from science in society to science for society*, in: *Science and Public Policy*, Vol. 39, No. 6, pp. 751-760.

Pagallo, Ugo (2017): *When morals ain't enough: Robots, ethics, and the rules of the law*, in: *Mind and Machines*, January 2017, Springer Online, pp. 1-14.

Pellerin, Cheryl (2015): "Work: Human-Machine Teaming Represents Defense Technology Future" *DoD News: Defense Media Activity,* November 8 2015, URL: https://www.defense.gov/News/Article/Article/628154/work-human-machine-teaming-represents-defense-technology-future/

Peterson, Andrea (2016): *In an apparent first, Dallas police used a robot to deliver bomb that killed shooting suspect*, in: *The Switch*, The Washington Post, July 8 2016, URL: https://www.washingtonpost.com/news/the-switch/wp/2016/07/08/dallas-police-used-a-robot-to-deliver-bomb-that-killed-shooting-suspect/?utm_term=.c0dc1bbbb354 [accessed March 21 2017].

Prigg, Mark (2014): *Who goes there? Samsung unveils robot sentry that can kill from two miles away*, in: *Mail Online*, September 15 2014, URL: http://www.dailymail.co.uk/sciencetech/article-2756847/Who-goes-Samsung-reveals-robot-sentry-set-eye-North-Korea.html [accessed March 21 2017].

Raytheon Corporation, Product Description: *Phalanx Close-In Weapons System*, URL: http://www.raytheon.com/capabilities/products/phalanx/ [accessed March 21 2017].

Rid, Thomas (2016): *Rise of the Machines – the Lost History of Cybernetics,* W.W. Norton & Company: New York.

Roberts, Jeff John (2016): *Why It's Legal for Police to Kill With a Robot*, in: *Fortune*; Tech, July 9 2016, URL: http://fortune.com/2016/07/09/robot-bomb [accessed March 21 2017].

Roff, Heather M. (2013*): Killing in War: Responsibility, Liability and Lethal Autonomous Robots*, in *Routledge Handbook of Ethics and War: Just War Theory in the 21st Century*, eds. Adam Henschke, Nick Evans and Fritz Allhoff (Routledge Press, 2013). **In this paper, the author argues that just war theory assumes the philosophical autonomy of the agents and patients involved in a lethal decision. Since Lethal Autonomous Robots (LAR) are autonomous only in the engineering sense, this means that Just War Theory and allied systems of thought are thus incapable of incorporating LARs into their ethical framework regarding the behavior of combatants.**

Roff, Heather M. (2014): *The Strategic Robot Problem: Lethal Autonomous Weapons in War*, in: *Journal of Military Ethics*, Vol. 13, No. 2, pp. 211-227. ***This paper argues that if one wants to understand the consequences of creating and deploying lethal autonomous weapons systems, one need to "look to the targeting process" which includes "how militaries actually create military objectives, and thus identify potential targets".***

Roff, Heather M. and Moyes, Richard (2016): *Meaningful Human Control, Artificial Intelligence and Autonomous Weapons*, Briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, 11-15 April 2016. ***This paper draws on thinking around 'meaningful human control' developed in collaboration with Dr. Heather Roff. With its name "Article 36" linked to the 1977 Additional Protocol I to the Geneva Conventions of August 1949 wherein states are required to review new weapons, this non-profit organization focuses on the prevention of "harm caused by existing weapons and to build a stronger framework to prevent harms as weapons are used or developed in the future".***

Royakkers, Lambèr and van Est, Rinie (2015): *A Literature Review on New Robotics: Automation from Love to War*, in: *International Journal of Social Robotics*, November 2015, Vol. 7, Issue 5, pp. 549-570.

Royakkers, Lambèr and van Est, Rinie (2016): *Just Ordinary Robots: Automation from Love to War*, CRC Press: Boca Raton.

Russell, Stuart J. and Norvig, Peter (2014): *Artificial intelligence: a modern approach*, Third Edition, Pearson Education: Harlow.

Russell, Stuart, et al. (2015): *Research Priorities for Robust and Beneficial Artificial Intelligence*, in: *Articles of the Association for the Advancement of Artificial Intelligence*, Palo Alto. See also: *Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter*, Future of Life Institute, URL: https://futureoflife.org/ai-open-letter/ [last accessed: 24.02.2017]. ***The Open Letter on Artificial Intelligence (2015) is signed by world-renowned scientists and artificial intelligence experts (i.e. Stephen Hawking, Peter Norvig, Eric Horvitz, Stuart Russell, Max Tegmark). Its intention is to warn from pitfalls of the technology and the danger to become an existential threat for humans, but also to set research priorities in order to avoid such outcomes and foster a "robust" artificial intelligence.***

Saxon, D. (2016): *Autonomous drones and individual criminal responsibility*, in: *Drones and responsibility: legal, philosophical and socio-technical perspectives*, Di Nucci & Santoni de Sio (eds.), pp. 148-166, Routledge: London.

Schmitt, Michael N. (2013): *Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics*, in: *Harvard National Security Journal Features*, Presidents and Fellows of Harvard College and Michael Schmidt.

Schmitt, Michael N. (ed.) (2013): *Tallinn Manual on the International Law Applicable to Cyber Warfare*, Cambridge University Press: Cambridge.

Schmitt, Michael N. and Thurnher, Jeffrey S. (2013): *'Out of the Loop': Autonomous Weapon Systems and the Law of Armed Conflict*, in: *Harvard National Security Journal, February 5 2013, Vol. 4 No. 231.* **The authors argue that a ban of Autonomous Weapon Systems would be "insupportable as a matter of law, policy, and operational good sense".**

Sharkey, Amanda (2014): *Robots and human dignity: A consideration of the effects of robot care on the dignity of older people*, in: *Ethics and Information Technology*, March 2014, Vol. 16, No. 1, pp. 63–75.

Sharkey, Noel (2007): *Automated Killers and the Computing Profession*, in: *Computer* (Journal), November 2007, Vol. 40, No. 11, pp. 122-124.

Sharkey, Noel (2009): *Death strikes from the sky*, in: *IEEE Technology and Society Magazine*, 28(1): 16–19.

Sharkey, Noel (2010): *Saying 'no!' to lethal autonomous targeting*, in: *Journal of Military Ethics* 9(4): 369–383.

Sharkey, Noel (2011): *Moral and legal aspects of military robots*, in: E*thical and Legal Aspects of Unmanned Systems*, Dabringer G (ed.), pp. 43–51, Institut für Religion und Frieden: Vienna.

Sharkey, Noel (2012a): *Killing Made Easy: From Joysticks to Politics*, in: *Robot Ethics: The Ethical and Social Implications of Robotics*, Lin, P., Abney, K., and Bekey, G.A., pp. 111-128, MIT Press: Cambridge.

Sharkey, Noel (2012b): *The evitability of autonomous robot warfare*, in: *International Review of the Red Cross*, Comments and Opinions, Summer 2012, Vol. 94, No. 886, pp. 787-799.

Sharkey, Noel (2016): *Staying in the loop: Human supervisory control of weapon*, in: *Autonomous Weapons Systems – Law Ethics Policy,* Bhuta N., Beck S., Geiss R., Liu H., Kress C. (eds.), Cambridge University Press: Cambridge, pp. 23-38.

Sharkey, Amanda and Sharkey, Noel (2012): *Granny and the robots: ethical issues in robot care for the elderly*, in *Ethics and Information Technology* 14 (1):27-40. **This paper reports on six ethical concerns raised by healthcare robots: (1) the potential reduction in the amount of human contact; (2) an increase in the feelings of objectification and loss of control; (3) a loss of privacy; (4) a loss of personal liberty; (5) deception and infantilization; (6) the circumstances in which elderly people should be allowed to control robots.**

Shladover, Steven E. (2016): The Truth about "Self-Driving" Cars, in: Scientific American, June 2016, 314(6): 52-57.

ShotSpotter: Law Enforcement, URL: http://shotspotter.com/law-enforcement [accessed June 16 2017].

Sidner, S. and Simon, M. (2016): *How Robot, Explosives Took out Dallas Sniper in Unprecedented Way*, in: *CNN Online*, July 12 2016, URL: http://edition.cnn.com/2016/07/12/us/dallas-police-robot-c4-explosives/ [accessed March 24 2017].

Singer, Peter W. (2009): *Wired for War - The Robotics Revolution and Conflict in the Twenty-first Century*, The Penguin Press: New York. ***This work is based on hundreds of interviews with people from the robotics field, politics, military, etc. It describes and approaches the subject of deployment and development of unmanned vehicles and robots in warfare critically.***

Snyder, Rosalyn G. (2001): *Robots assist in search and rescue efforts at wtc*, in: *IEEE Robotics and Automation Magazine, Vol. 8, No. 4, p.26-28.*

Sorell, Tom and Draper, Heather (2014): *Robot carers, ethics, and older people*, in: *Ethics and Information Technology*, September 2014, Vol. 16, No. 3, pp. 183-195.

Sotala, K. and Yampolskiy R. (2013): *Responses to Catastrophic AGI Risk: A Survey*. Technical report 2013-2. Berkeley, CA: Machine Intelligence Research Institute.

Sparrow, Robert (2007): *Killer Robots*, in: *Journal of Applied Philosophy*, Vol. 24, No. 1, pp. 62-77.

Sparrow, Robert (2009a): *Building a better warbot: Ethical issues in the design of unmanned systems for military applications*, in *Science and Engineering Ethics*. 15, 2, p. 169 – 187.

Sparrow, Robert (2009b): *Predators or plowshares? Arms control of robotic weapons*, In *IEEE Technology and Society Magazine*. 28, 1, p. 25 – 29.

Sparrow, Robert (2011): *Robotic weapons and the future of war*, in *New Wars and New Soldiers: Military Ethics in the Contemporary World*. Tripodi, P. & Wolfendale, J. (eds.). Surrey, UK: Ashgate Publishing Limited, p. 117 - 133.

Sparrow, Robert (2013): *War without Virtue?*, in: *Killing by Remote Control*, Strawser, B. J. (ed.), pp. 94-105, Oxford University Press: New York.

Steinhoff, Uwe (2013): *Killing Them Softly: Extreme Asymmetry and its Discontents*, in: *Killing by Remote Control*, Strawser, B. J. (ed.), pp. 179-210, Oxford University Press: New York.

Strawser, Bradley J. (2010): *Moral predators: The duty to employ uninhabited aerial vehicles*, in: *Journal of Military Ethics,* Vol. 9, No. 4, pp. 342–368.

Strawser, Bradley J. (2013): *Killing by remote control: The Ethics of an unmanned military*, Oxford University Press: Oxford.

Sullins, John P. (2012): *Robots, Love, and Sex: The Ethics of Building a Love Machine*, in: *IEEE Transactions on affective Computing,* Vol. 3, No. 4, October-December 2012.

Sullins, John. P. (2010): *RoboWarfare: Can Robots be More Ethical Than Humans on the Battlefield?*, in: *Ethics and Information technology*, Vol. 12 No. 3, pp. 263-275.

Sullins, John P. (2006): "When Is a Robot a Moral Agent?" International Review of Information Ethics, Vol. 6, No.12, pp. 23-30: December. **This paper argues for the minimum requirements for artificial moral agency in robotics**.

Thurnher, Jeffrey S. (2013): *The Law That Applies to Autonomous Weapon Systems*, in: *insights*, American Society of International Law, Vol. 17, No. 4, January 18 2013. ***This immediate reaction to the report "Losing Humanity" (see: Docherty, Bonnie et al.: 2012) tries to make clear from a legal perspective that it is important to distinguish "policy, morality or ethical arguments" from "purely legal ones" in regards to the field of international law.***

Turkle, Sherry (2011): *Alone Together: Why We Expect More from Technology and Less from Each Other*, Basic Books: New York.

Turkle, Sherry (2015): *Reclaiming Conversation: The Power of Talk in a Digital Age*, Penguin Press: New York.

Turse, Nick and Engelhardt, Tom (2012): *Terminator Planet: The First History of Drone Warfare, 2001-2050*, Dispatch Books/CreateSpace Independent Publishing Platform.

United Nations (1990): *Basic Principles on the Use of Force and Firearms by Law Enforcement Officials* (September 7 1990), Adopted by the Eighth United Nations Congress on the Prevention of Crime and the Treatment of Offenders, August 27-September 7 1990: Havana.

United Nations (2013): *Report CCW/MSP/2013/10* (December 16 2013), Meeting of the High Contracting Parties to the 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, November 14-15 2013: Geneva.

United Nations (2014): *Report CCW/MSP/2014/9* (November 27 2014), Meeting of the High Contracting Parties to the 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, 13-14 November 2014: Geneva. Available at: http://www.unog.ch/80256EE600585943/(httpPages)/A038DEA1DA906F9DC1257DD90042E261?OpenDocument

United Nations (2015): *CCW Meeting of the High Contracting Parties*, Final Report, Advanced Version, November 12-13 2015: Geneva. Available at: http://www.unog.ch/80256EE600585943/(httpPages)/6CE049BE22EC75A2C1257C8D00513E26?OpenDocument

United Nations (2016): *Report CCW/CONF.V/10* (December 23 3016), Advance Version, Fifth Review Conference of the High Contracting Parties to the 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, December 12-16 2016: Geneva.

United Nations (2016b): *Recommendations to the 2016 Review Conference*, Advanced Version, 2016 Meeting of Experts on LAWS, Submitted by the Chairperson of the Informal Meeting, URL: http://www.unog.ch/80256EDD006B8954/(httpAssets)/6BB8A498B0A12A03C1257FDB00382863/$file/Recommendations_LAWS_2016_AdvancedVersion+(4+paras)+.pdf [accessed March 31 2017]. See further details: Reaching Critical Will Organisation (2016): CCW Report, April 15 2016, Vol. 3, No. 5, URL: http://www.reachingcriticalwill.org/disarmament-fora/ccw/2016/laws/ccwreport [accessed April 15 2017].

United States Army (Dec, 2006)**:** *Field Manual 3-24*, *Counterinsurgency (COIN-FM3-24),* December 2016, Department of the Army: Washington.

US Department of Defense (2012): *Autonomy in Weapon Systems*, Directive No. DoDD 3000.09, November 21 2012.

US Department of Defense (2015): Law of War Manual. Available at: https://publicintelligence.net/dod-law-of-war/ [accessed July 16 2017]

Vallabhaneni, Anirudh et al. (2005): *Brain–computer interface*, in: *Neural Engineering*, He Bin (ed.), pp. 85-121, Kluwer Academic/Plenum Publishers: New York.

Vallor, Shannon (2016): *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*, Oxford University Press: New York.

van den Brule, Rik et al. (2014): *Do Robot Performance and Behavioral Style affect Human Trust?*, in: *International Journal of Social Robotics*, November 2014, Vol. 6, No. 4, pp. 519-531.

van den Hoven, Jeroen (2007): *ICT and Value Sensitive Design*, in: *The Information Society: Innovation, Legitimacy, Ethics and Democracy*, Goujon P., Lavelle S., Duquenoy P., Kimppa K., Laurent V. (eds.), IFIP International Federation for Information Processing, Vol. 233, pp. 67-72Springer: Boston MA.

van den Hoven, Jeroen (2008): *Information Technology and moral philosophy*, Cambridge University Press: Cambridge.

van den Hoven, Jeroen (2014): *Responsible innovation: a new look at technology and ethics*, in: *Responsible innovation 1: innovative solutions for global issues*, Van den Hoven et al. (eds.), pp. 3-13, Springer: Dordrecht.

van der Vyver, J.-J. et al. (2004): *Towards genuine machine autonomy*, in: *Robotics and Autonomous Systems*, Vol. 46, No. 3, pp. 151-157.

van Wynsberghe, Aiimee (2015): *Healthcare Robotics: Ethics, Design and Implementation*, Routledge: London.

van Wynsberghe, Aiimee and Nagenborg, Michael (2016): *Civilizing drones by design*, in: *Drones and responsibility: legal, philosophical and socio-technical perspectives*, Di Nucci & Santoni de Sio (eds.), pp. 148-166, Routledge: London.

Veruggio, Gianmarco and Operto, Fiorella (2008): *Roboethics: Social and ethical implications of robotics*, in: *Springer Handbook of Robotics*, Siciliano B. and Khatib, O. (eds.), pp. 1499-1524, Springer: Berlin.

von Bothmer, Fredrik (2014): *Robots in Court - Responsibility for Lethal Autonomous Weapons Systems*, in: *Mensch und Maschine - Symbiose oder Parasitismus?*, pp. 102-112, Brändli, Sandra et al. (ed.), Schriften der Assistierenden der Universität St. Gallen (HSG), Vol. 9, Stämpfli Verlag: Bern.

Wallach, Wendell (2013): *Terminating the Terminator*, in: *Science Progress*, Center for American Progress, January 29 2013, Online Journal, URL: https://scienceprogress.org/2013/01/terminating-the-terminator-what-to-do-about-autonomous-weapons/ [accessed March 28 2017].

Wallach, Wendell and Allen, Colin (2013), *Framing Robot Arms Control*, in: *Ethics and Information Technology*, Vol. 15, No. 2, pp.125-135, Springer: Dordrecht. ***In this article, the authors apply concepts from their work on autonomous robots in general to the task of autonomous weapons arms control.***

Wallach, Wendell and Allen, Colin (2009): *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press: Oxford. ***As the book title indicates: Wallach and Allen aim to implement morality into robotic systems, namely to create "artificial moral agents (AMAs)", so that robots can i.e. distinguish "good" from "bad". They tackle this large-scale idea from an ethical, engineering and cognitive science perspective. As more as artificial agents are engaged in responsible decisions within their sphere of action, as more it becomes inevitable from Wallach and Allen's perspective to develop and implement solutions that allow robots to make moral decisions - whether such a morality is directly implemented or self-learned by robotic systems.***

Walter, Christian (2015): *Cyber Security als Herausforderung für das Völkerrecht*, in: *JuristenZeitung*, Vol. 70, No. 14, July 2015, pp. 685-693(9).

Weaver, John Frank (2014): *Robots are People too – How Siri, Google Car, and Artificial Intelligence will Force Us to Change Our Laws*, Praeger: Santa Barbara.

Weng, Yueh-Hsuan et al. (2015): *Intersection of "Tokku" Special Zone, Robots, and the Law: A Case Study on Legal Impacts to Humanoid Robots*, in: *International Journal of Social Robotics*, Vol. 7, No. 5, pp. 841-857.

Winfield, Alain (2012): *Robotics: A Very Short Introduction*, Oxford University Press: Oxford.

World Economic Forum (WEF) (2016): *Robots in war: the next weapons of mass destruction?*, 17. January 2016, URL: https://www.weforum.org/agenda/2016/01/robots-in-war-the-next-weapons-of-mass-destruction/ [accessed October 12 2017].