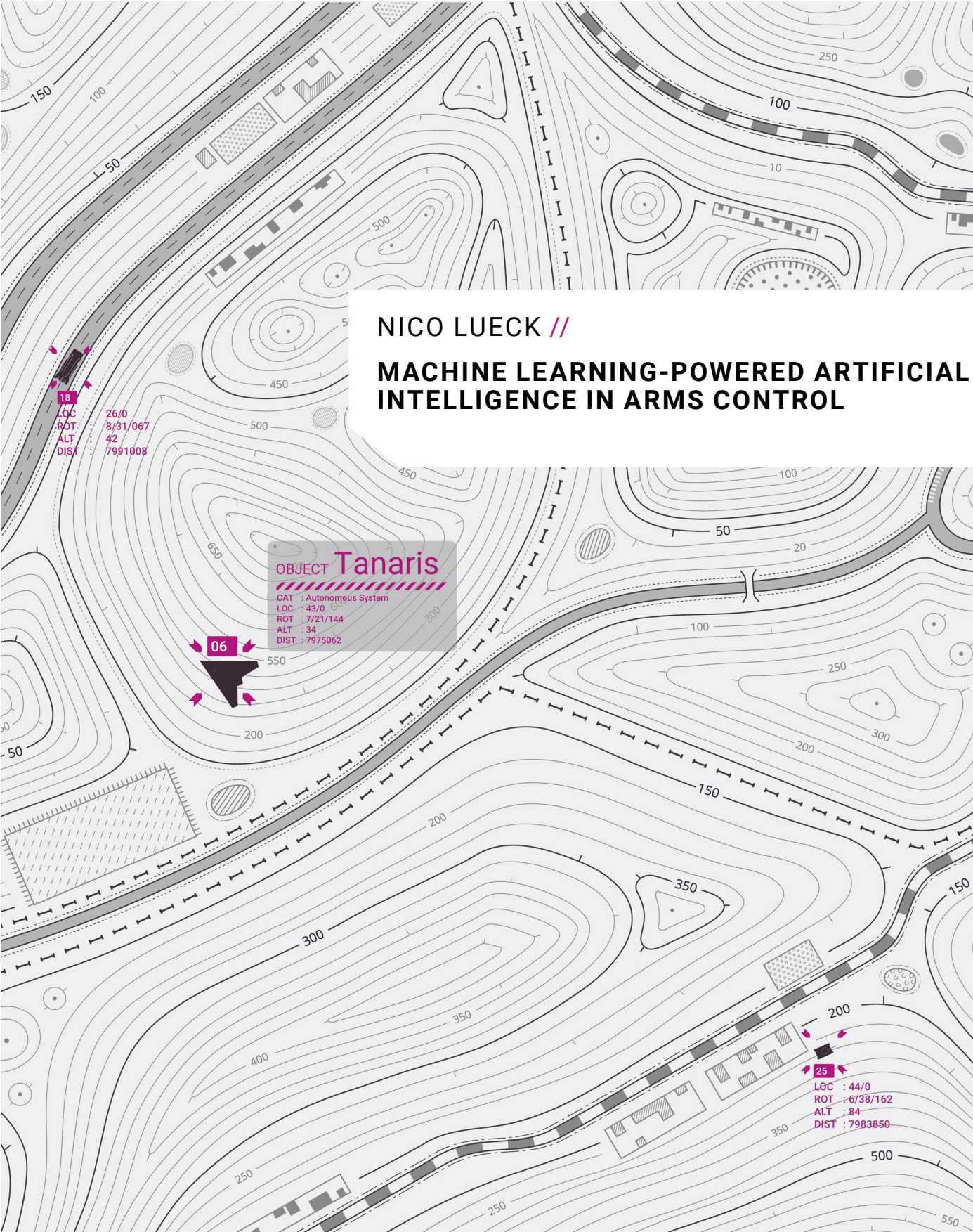


PRIF REPORT

PEACE RESEARCH INSTITUTE FRANKFURT / LEIBNIZ-INSTITUT HESSISCHE STIFTUNG FRIEDENS- UND KONFLIKTFORSCHUNG



PRIF Report 8/2019

MACHINE LEARNING-POWERED ARTIFICIAL INTELLIGENCE IN ARMS CONTROL

NICO LUECK //

LEIBNIZ-INSTITUT HESSISCHE STIFTUNG FRIEDENS- UND KONFLIKTFORSCHUNG (HSFK)
PEACE RESEARCH INSTITUTE FRANKFURT (PRIF)

Cover:

Nico Lueck, own figure

Text license:

Creative Commons CC-BY-ND (Attribution/NoDerivatives/4.0 International).

The images used are subject to their own licenses.



Correspondence to:

Peace Research Institute Frankfurt

Baseler Straße 27–31

D-60329 Frankfurt am Main

Telephone: +49 69 95 91 04-0

E-Mail: info@hsfk.de

<https://www.prif.org>

ISBN: 978-3-946459-51-4

Everybody is talking about artificial intelligence (AI) and the subject is considered one of the central issues of the future, because it is foreseeable that the power of AI systems will enable extraordinary advances in a wide variety of applications – resource optimization, forecasting, object recognition, human-computer interaction or controlling robot-based systems. This is even more the case when AI possesses the capacity for so-called “machine learning”, i.e., the ability to develop its own rules on the basis of observations or provided data. This capability allows unprecedented independence from the capability of human programmers to anticipate events.

In armaments, these developments are already playing an important role: Conventional AI is already being used in combat aircraft, drones, gun turrets or humanoid robots as a control and assistance unit, for example in navigation and target recognition. Machine learning-powered Artificial Intelligence (MLpAI) is currently being tested in the development of new weapon systems or integrated into prototypes.

These trends make artificial intelligence an important issue in arms control, and in two respects. As the object of arms control, MLpAI lies outside traditional approaches because it has neither the physical qualities or abilities nor the transparent operations that current methods and procedures for both quantitative and qualitative arms limitation are based on. On the other hand, MLpAI provides new tools for arms control. Thus, it is conceivable that the verification of existing and new arms control contracts, i.e., verification of compliance with them, could benefit significantly from MLpAI as a technical tool, for example by increasing the precision and speed of collecting, processing and analyzing data.

From a security policy perspective, MLpAI thus offers risks and opportunities alike, and the foreseeable increase in the use of machine learning will increase these risks and opportunities to a very considerable degree. The risk is that MLpAI, as the core element of future autonomous weapons systems, must be limited by arms control, but that at the same time arms control does not possess adequate technical capabilities. At this point, traditional approaches have been exhausted and the possibility of monitoring and limiting MLpAI during development or deployment remain. If MLpAI is used unchecked, it jeopardizes strategic stability by minimizing the de-escalating human factor, promoting a technological arms race, and spreading uncontrollably. This is countered by the enormous potential inherent in MLpAI for the verification of arms control agreements. The identification of objects, phenomena and changes over time on satellite imagery, on video recordings or in data from electromagnetic, seismic or acoustic sensors is demonstrably improved by the use of machine learning. Far more accurate and comprehensive information processing could increase transparency, discourage actors from violating agreements or be used as evidence of compliance with agreements.

The report concludes that MLpAI is both part of the problem and part of the solution. Current and future application examples show that MLpAI is the core element of modern weapons systems and thus should itself be the subject of arms control – especially given that the capabilities, but also the dangers, of the weapons systems would be greatly increased by MLpAI. However, as an object of control it eludes many approaches to qualitative or quantitative limitation. Thus, it increases the relevance of alternative methods for overall military transparency and confidence-building. It is precisely

in these methods, however, that it can be used as a verification instrument. Through precise and extensive information processing it can create greater transparency and verify compliance with agreements, thus bolstering trust between parties. The developments in the two possible areas where MLpAI can be deployed and the lower technical complexity in verification measures show that early action can support the potential benefits of MLpAI, because it can now help arms control develop new capacity before it faces the new challenge of MLpAI-controlled weapons systems.

1. Introduction	1
2. Machine Learning-powered Artificial Intelligence	2
2.1 Driver of the AI revolution: Machine Learning	3
2.2 Challenges in the development and use of Machine Learning	5
3. Machine Learning-powered Artificial Intelligence in Weapons Systems	7
3.1 AI as a core element of modern weapons systems and learning ability as a multiplier	8
3.2 Characteristics making control difficult and new approaches	9
3.3 Conceptual reflection: Strategic instability and proliferation	12
4. Machine Learning-powered Artificial Intelligence in Verification Measures	16
4.1 Applications and technical potential	16
4.2 Conceptual reflection: Improvements in existing methods	22
5. Machine Learning-powered Artificial Intelligence is part of the problem and part of the solution	23
References	24

1. INTRODUCTION

The application of artificial intelligence (AI) in weapons systems without external monitoring or controls poses a risk to humanity that is increasing with the growing capability of the technology. The risk is most obvious when AI is intentionally applied for destructive purposes or when it chooses such a path on its own initiative, in order to achieve a given goal. On the other hand, the development of AI also offers positive opportunities for humanity. The extraordinary power of AI to process information enables it to recognize and interpret patterns in seemingly unstructured datasets, solve problems based on pre-specified or acquired knowledge, plan measures, or draw conclusions. In practice, AI is used in areas such as optimizing the use of resources, forecasting developments over time, recognizing objects in recorded images, communicating with humans or controlling robotic systems. The tension between risks and opportunities in the use of AI is thus the basic issue examined in this report.

The ambivalence between positive and negative consequences of the implementation of AI is also found in arms control: The overarching objective of arms control is to “create a greater measure of strategic stability between two or more states” and thus “reduce the likelihood of war breaking out at a time of crisis” (Croft 1996: 91–92). For this purpose, arms control measures “(a) freeze, limit, reduce or abolish certain categories of weapons; (b) ban the testing of certain weapons; (c) prevent certain military activities; (d) regulate the deployment of armed forces” (Goldblat 2002: 3).

As with any innovations in weapons technologies, governments seek to use technology in weapons systems to gain or offset technological superiority and thus obtain a strategic advantage over other states. Consequently, if the use of AI in weapons systems were regulated in order to improve strategic stability, this would be a new challenge for arms control. As an object of control, AI would be part of a lineup including mines, ammunition, small arms, conventional weapons, weapons of mass destruction and delivery systems. However, these are physical objects that are regulated by technical, geographical or application-related characteristics. If this is also to be done with AI, inherent properties of AI must be identified that make restriction and monitoring possible. On the other hand, if AI is not regulated, previously unheard-of consequences for strategic stability may arise.

This threat to stability is emerging at a time when arms control is already in crisis due to military-technical progress, breaches of treaties and lack of political will (Arbatov 2015; Schmidt 2017). But at the same time, AI provides new opportunities for controlling conventional weapons and weapons of mass destruction and can help verify compliance with existing and new arms control treaties. In order to increase confidence in control treaties and among states, off-site and on-site inspections create transparency and help to monitor states’ performance. For this purpose, technical aids are usually deployed to collect, process and analyze information (Goldblat 2002: 310). The potential of verification measures has increased with new technologies as satellites, sensors and other monitoring techniques improve the information situation (Pilat 2002: 81). At this point AI can serve as a multiplier because it can analyze this information – especially large data sets – with greater precision and at higher speed than conventional methods.

This report first discusses the research object “Machine Learning-powered Artificial Intelligence” (MLpAI) and current challenges in developing the ability to learn. It then explores the question of the threats arising from application of MLpAI in weapons systems (Chapter 3) and where it opens up new opportunities for arms control and verification (Chapter 4). The report concludes that MLpAI in weapons systems can avoid control by many traditional arms control approaches, and at the same time can increase fundamental transparency and reinforce trust among parties to verification measures (Chapter 5).

2. MACHINE LEARNING-POWERED ARTIFICIAL INTELLIGENCE

When an attempt is made to answer the question of what is meant by “AI” it must be admitted that there are many approaches to the issue, but no universally accepted definition. “There are about as many definitions of AI as researchers developing the technology” (McCloskey 2017).

But most definitions agree on two core characteristics: 1) solving highly complex tasks and 2) adapting to the environment.¹ AI systems that perform especially well in these two core areas mostly display the ability to learn. This feature stands out among the recognized abilities of AI – perception, knowledge representation, problem solving, planning and reasoning. This is supported by the still ongoing increases in performance of hardware, especially the computing power of the processors, and the availability of large data sets as a learning tool.

In the past, computer programs could interpret only matters for which the programmer had specified rules, in other words, conditional “if-then” relationships. When future situations or changes over time cannot be anticipated by the human programmer, or if the programmers themselves do not know the solution, then the use of machine learning helps (Russell/Norvig 2010: 693). This refers to systems that generate their own knowledge by extracting patterns from raw data. With this innovation, AI can solve complex real-world problems without having to rely on solutions specified by programmers (Goodfellow 2016: 3). The focus of this report will be on such machine learning-powered artificial intelligence (MLpAI), because the ability to learn is a prerequisite for outstanding achievements in the two core properties of AI mentioned above – the ability to solve complex tasks and adaptability.

In addition to its deliberate focus on specific learning ability, this report also restricts the aspects of AI to be examined according to their general goal: Is the goal a generally intelligent machine or a machine that is considered “intelligent” only in a specific discipline? Essentially, every definition can be classified according to this division. Although the vision of a generally intelligent machine arouses public interest, there is no system yet that satisfies this requirement. Since experts estimate the development time for so-called Artificial General Intelligence (AGI) to be at least 50 years (Müller/Bostrom 2016: 559), such a definition is not a practical starting point for analyzing current and recent applications. Instead, in this report exclusively examples of AI will be examined which are being

1 A summary of definitions of AI can be found in Artificial General Intelligence Sentinel Initiative (2017); Legg/Hutter (2007).

developed or optimized for a specific discipline or task. For this already existing type of AI human beings do not serve as a model, but at the most as a performance criterion. The goal of “application-specific” AI – also referred to as “*narrow AI*” (Franklin 2014: 16) or “*weak AI*” (Searle 1980: 417) – is to achieve worthwhile results or “intelligent” performance in at least a single discipline.

By focusing on application-specific machine learning-powered artificial intelligence, the review of empirical findings in this report focuses primarily on new, innovative AI programs and excludes both traditional computer programs and futuristic AI concepts. This report makes a deliberate distinction between “AI” and “MLpAI”. Where the term “AI” is referred to on its own, then either the statements are also valid for AI without learning capability or the functioning of an empirical example cannot be attributed with certainty to the ability to learn.

2.1 DRIVER OF THE AI REVOLUTION: MACHINE LEARNING

Many human or animal actions are highly complex for computers. The processing of speech or images and subsequent coordination of actions is so complex that human beings themselves do not understand it sufficiently to be able to transfer it to a program. In order for these highly complex tasks to be mastered by AI, it is given the ability to learn. MLpAI recognizes solution patterns and transfers them to other tasks. This is also necessary because tasks may change depending on the time, the user or other parameters (Shalev-Shwartz/Ben-David 2014: 21–22).

Two forms of MLpAI can be distinguished: AI can learn purely on the basis of pre-determined data (*unsupervised*) or with the help of additional inputs from a teacher (*supervised* or *reinforced*).

Unsupervised learning occurs when machine learning builds clusters from unknown data. Machine learning independently identifies characteristics that involve similarities and differences. For example, it recognizes that satellite images can be grouped according to visible land, sea or clouds. In the case of supervised or reinforced learning, on the other hand, the relevant characteristics of the data are pre-defined. This approach allows classification (e.g., mapping subjects into images), regression (e.g., prediction of environmental phenomena), or detection of anomalies (e.g., detection of unnatural eruptions) within the data (Russell/Norvig 2010: 694–697).

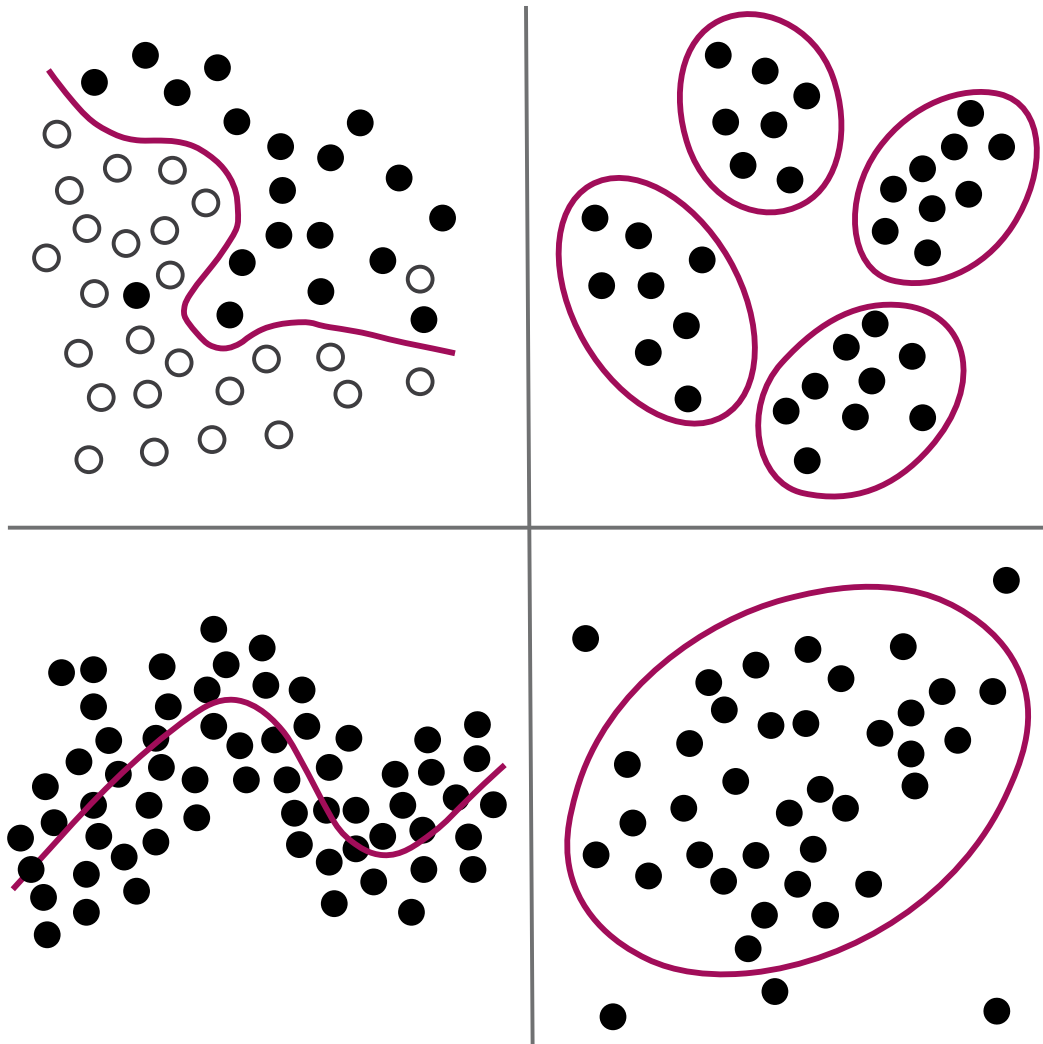


Fig. 1: Classification (top left), cluster building (top right), regression (bottom left), recognition of anomalies (bottom right). Author's own depiction.

Machine learning can involve numerous methods such as *naive Bayesian algorithms*, *support vector machine algorithms*² or different variants of decision trees. However, the current method of *deep learning* (also known as the *deep neural network* method) accounts for most of the progress made in AI research recently. Its high level of adaptability and performance in highly complex tasks is based on its orientation to the human brain by copying neural networks. But so far it has not even

² These methods are based on mathematical methods that maximize different variables in functions in order to classify objects, for example, to maximize the difference between objects or achieve the lowest classification costs for a predefined measure.

approached the complexity of the human brain (Hawkins/Blakeslee 2004: 25–27). However, the improvements achieved in performance in specific applications such as image recognition are already substantial. While conventional programs are unable to extract meaning from a collection of different pixels of an image, the *deep learning* method divides the process into many sub-steps. The first level (“1st hidden layer”) compares the brightness of the pixels, in order to identify edges in the image. The second level searches for corners and contours based on the results of the first level. The third level, in turn, can then identify entire parts of specific objects (e.g., nose, ears, legs) by finding specific sub-groupings of corners and contours. Finally, the computer is able to interpret the image on the basis of the various objects detected (Goodfellow 2016: 6).

This process can only be performed after the computer has previously learned which objects an image includes, for example, a car, a person, or an animal. In the learning process, the “output layer” is specified in advance and the levels are processed in reverse order. Machine learning associates the objects detected at the third level with the given result, the contours of the second level with the objects, and the brightness of the pixels of the first level with the contours.³ It is extremely important to understand that the results at which the *deep learning* program arrives only reflect probabilities. It does not decide on a particular option and justify it, but, for example, states that it is 9% likely to be a car, 72% a person or 19% an animal. How these outputs are applied is up to the programmer.

Because machines understand only the formalized interdependencies of variables and not their content, machine learning is called “automated statistics” by some scientists and critics (Danks 2014: 159f). However, the learning models themselves and the parent program are based on a rule-based system. Thus – in contrast to statistics – they always contain conscious or unconscious normative ideas, which are introduced by the programmers via decisions in the program code or the selection of learning data (Algorithm Watch 2017: 3).

It should be noted that the potential of machine learning is far from exhausted. Numerous methods covered by the term have already been and are still being developed (Farrelly 2016). In addition to increased complexity, future milestones are expected in the following areas: continued learning beyond a limited set of input data, transfer of acquired knowledge to other tasks, independent generation of input data (Morisse 2017) and learning by observing other machines (Li et al. 2016).

2.2 CHALLENGES IN THE DEVELOPMENT AND USE OF MACHINE LEARNING

If machine learning is to be used in real applications, three challenges must be overcome: (1) errors or inaccuracies in the system or learning process, (2) difficult explainability⁴ of decisions in the absence of appropriate precautions, and (3) manipulability of results by means of deliberately biased input data.

3 A more detailed explanation and excellent visualization of the process can be found in Zeiler and Fergus (2013).

4 The degree to which learned models and decisions can be understood and trusted by users.

The first challenge for machine learning presented here is to strike a balance between two factors, namely, on the one hand, the complexity of the learned model, which becomes more complex with increasing precision, and, on the other hand, the generalizability of the learned model to new data (Danks 2014: 155f). The interaction between these factors poses a dilemma: If the model becomes too complex and exactly maps the training data – i.e., the data which is supposed to describe reality – then the model can no longer generalize, so it cannot be applied to new, previously unknown data (*overmatching*). If the model maps the training data too inaccurately, then it can be generalized, but reality is not correctly modeled (*undermatching*). As a result, generalization is vulnerable to incorrect conclusions. Harmful feedback loops can occur in which machine learning designs a logic of its own, from which it cannot escape without corrective feedback. These generalization errors are not always detected in practical situations and may thus be adopted in real-life applications. The actors who depend on this incorrect model of AI must follow the rules established by the AI to succeed. This, in turn, distorts possible feedback to the system. This problem can be seen for example in the prediction of criminal offenses: Police officers patrol separately in districts that the system rates as particularly critical at the current time of day. Because of the increased police presence more offenses are detected. The prediction made by the system is confirmed and at the same time the incidents that are logged are included in the data base for further predictions. Such feedback loops have also been detected in facial recognition, teacher assessment, and granting of credit and insurance coverage (O’Neil 2016).

As a second challenge, it is particularly noteworthy that machine learning, especially *deep learning* has been described by many researchers as a “black box,” because of its inherent characteristics (Knight 2017; Ribeiro et al. 2016). Given that in their entirety autonomously learned rules are a mathematical model based on thousands or even millions of parameters, a decision path in complex models involving, for example, the assignment of a small section of an image to a category, cannot be logically understood by a human

Even the developers cannot always understand how machine learning arrives at its decision. In many cases this uncertainty is accepted or simply left up to the system. But with increasing application of the method of *deep learning* this becomes an increasing risk. Although other methods are more explainable, explainability correlates negatively with accuracy. For example, if a program learns by using a decision tree with binary branches, the results are transparent, but not as accurate as through the simultaneous, percentage weighting of several features at hundreds of abstraction levels provided by a *deep learning* model (Gunning 2016: 4). In the case of *deep learning* systems, an attempt is made to increase explainability by not only specifying the probability of a particular option, but also giving a justification of why this option is the most probable. For example, instead of the output, “the object is 93% likely to be a cat,” the machine learning algorithm should specify: “The object is 93% likely to be a cat, because it has fur, paws and claws”. Research on linking pattern recognition and verbal description of the pattern shows that additional machine learning applications can make such explanations possible (Park et al. 2017). Although some approaches may sound promising, this branch of research⁵ is still new and restricted in scope, at least in terms of what is publicly known.

5 This research is referred to as “explainable artificial intelligence.”

By contrast, US military research has dedicated a program area of its own to explainable machine learning (Gunning 2016).

The third challenge became visible as a result of the insights of a research team from Google, Facebook and various universities. The team found that the image recognition of the *deep learning* method had an unexpectedly high error rate and contained errors of classification if information that is not visible to the human eye had been added to the image (Szegedy et al. 2014). These results lead to two fundamental findings: *First*, even machine learning with outstanding performance capability fails to learn the correct concept underlying the images. Instead, it constructs a model based on statistical relationships that incorporates naturally occurring data, it is true, but has fundamental weaknesses when confronted with a very unnatural or unlikely data distribution (Goodfellow et al. 2015: 2). *Secondly*, the possibility of manipulating machine learning via input data creates a major gap in security. Without a solution there is a constant risk that the data the AI analyzes has been deliberately altered in such a way that the system misinterprets it. In addition, scientists have been able to show that a conflicting image (*adversarial example*⁶) retains its manipulative properties even when it is printed and re-photographed with any camera. For example, a picture of a washing machine may look like a safe to the algorithm (Kurakin et al. 2017). In another study it was shown that facial recognition systems could be influenced with special glasses in such a way that the people photographed were identified as other people in the database. The authors warn that such methods could be used in future in criminal files (Sharif et al. 2016). This optical illusion for machines bears enormous risks for practical applications of machine learning. If attackers wanted to manipulate a self-driving car, they could alter street signs in such a way that the machine learning algorithm, which recognizes and interprets traffic signs, would interpret a sign as a right-of-way signal instead of a stop sign. That this is not just a hypothetical scenario was proved by Nicolas Papernot and his colleagues, who reconstructed the scenario just described (Papernot et al. 2017). These cases make it clear that manipulations are possible not only by changes to the data file, but also through purely visual changes, as objects in the real world can be manipulated too. In addition, the learning models involved in recognition of the washing machine, faces, and traffic signs were correctly trained, but the application data – i.e., the data to be analyzed – was manipulated.

3. MACHINE LEARNING-POWERED ARTIFICIAL INTELLIGENCE IN WEAPONS SYSTEMS

Under the aegis of the UN, a debate on the prohibition of deadly autonomous weapons systems is currently taking place in Geneva (Boulain/Verbruggen 2017). In civil society, calls for the banning of autonomous weapons systems are being supported by prominent scientists and organizations (Sauer 2016; Future of Life Institute 2015; Human Rights Watch 2012). While these debates and calls also address the use of AI in weapons systems, they focus on the issue of autonomy. As will be shown in this chapter, as a property of AI-controlled systems autonomy is not a suitable approach for arms control. Since other approaches of traditional arms control (physical characteristics or capa-

6 An adversarial example is input data intentionally designed to cause the system to crash or make a mistake.

bilities as well as internal functioning) do not provide reliable evidence in the case of AI, alternative control methods focused on development and deployment are considered. Considering the potential use of AI in weapons systems, destabilizing consequences such as accelerated processes, lack of de-escalating action, a technological arms race or uncontrolled proliferation are discussed.

3.1 AI AS A CORE ELEMENT OF MODERN WEAPONS SYSTEMS AND LEARNING ABILITY AS A MULTIPLIER

As with conventional computers, weapons systems also distinguish between hardware and software. This distinction is reflected in the terms “robotics” and “AI.” Developments in robotics are improving the physical capability and firepower of weapons systems, but their potential for improvement is limited by restrictions deriving from the laws of physics. The development of MLpAI, on the other hand, multiplies the software-related capabilities of weapons systems many times over, because only MLpAI can increasingly improve the ability of weapon systems to operate in complex environments.

Even if no general definition of autonomous weapons systems yet exists, application-specific AI is already to be found in weapons systems as a control or assistance unit. The ability to learn cannot be inferred with certainty in the following examples, but – depending on the system – AI is used for navigation, target recognition and identification, as well as in attack planning and execution. Many applications are located in digital space, in airspace, and in static defense systems, because the environment in which AI must operate in these applications is less complex than in ground or urban warfare. The environment increases in complexity as the number of (unknown) challenges increases: e.g., navigation on uneven ground, obstacles of all kinds and interactions with unfamiliar objects.

As a support system for fighter pilots, AI is trained, for example, to recognize targets based on radar images in order to avoid misjudgments by the pilot, or to make a decision to fire from a great distance significantly beyond the visual range of the pilot (Keller 2015). AI takes over further tasks in the drone *Taranis*, which British manufacturer *BAE Systems* is currently developing. In addition to manual remote control and automatic flight navigation, *Taranis* has a mode in which it autonomously plans a route and searches for targets until it reaches its mission target (Stevenson 2016). The AI *ALPHA*, which has not yet been installed in drones, has the capability to take command of the entire flight and battle maneuvers – until recently still the exclusive domain of human pilots: In a simulation opposing the experienced Colonel of the *US Air Force*, Gene Lee, the system demonstrated outstanding capability, simultaneously dodging missiles fired at it, firing at several targets, participating in coordinated maneuvers, and registering and learning from enemy tactics. The colonel labelled the MLpAI, which ran on a computer costing only \$US35.00, a *Raspberry Pi*, as “the most aggressive, responsive, dynamic and credible AI I’ve seen to date” (Ernest et al. 2016).

In addition, some static defense systems – small turrets or anti-aircraft guns – are among the first AI-controlled weapons systems, because they do not encounter complex environmental challenges. The *Super Aegis II* gun turret is designed to independently identify, target, track and ultimately fire upon targets without human help. Due to the fear of customers that the system could make mistakes in autonomous mode, the degree of autonomy can now be set individually (Parkin 2015).

The US Navy's *Phalanx CIWS* anti-aircraft cannon can also carry out these actions autonomously to defend itself against incoming missiles and aircraft (US Navy 2017).

The areas of application are constantly being expanded and autonomous nanodrones (Daniels 2017), warships (Courtland 2016), and humanoid robots (Boston Dynamics 2018) are already under development. However, development is not focused exclusively on individual systems, but also on a new form of interaction. In the future, weapons systems will also be able to operate in a swarm. A swarm consists of many individual machines that can act independently, but also in concert. They coordinate themselves based on uninterrupted communication among units (Ben-Ari/Mondada 2018: 251–252). This has the advantage, among other things, that there is no central control unit that can fail, so that individual defects or kills have only a small effect on the performance of the swarm. AI would thus not only lead to the technical superiority of individual systems but would also optimize the deployment of entire battle units.

No operational weapons system which uses the ability of machine learning is yet known. The *ALPHA* AI just described for controlling a fighter jet and the announcement by the Russian arms supplier *Kalashnikov* that it is utilizing learning capability (Russia Today 2017) seem to show that AI capable of learning is being integrated into new weapons systems. Although its operation is unknown due to military secrecy, private-sector research projects indicate how MLpAI could be used in weapons systems. A research team from chipmaker Nvidia, which actually develops highly specialized graphics chips, trained MLpAI to drive a car without dictating any rules. As input data, the MLpAI only received the movements of the steering wheel and camera footage from the front of the car. Despite this limited perception, the MLpAI learned the rules of road traffic in the course of human-controlled journeys and was subsequently able to drive on its own (Bojarski et al. 2016). This differs greatly from conventional autonomous driving systems, which are given the interpretation of traffic rules and vehicle behavior in advance. The learned driving style has the obvious disadvantage that human errors are acquired as well. However, it also has great advantages because the program learns intuitive rules about which the driver is not consciously aware and is also trained to deal with unanticipated situations. Accordingly, in an analogous way, mobile weapon systems could acquire the ability to navigate through supervised learning. Targeted recognition and identification can also be significantly refined through machine learning by using the typical skills of the *deep learning* method – perception and classification of objects. In (combat) situations AI has to decide and act appropriately. The necessary inference and planning skills can be trained by using simulations that are monitored by humans. Thus, machine learning can improve the performance of all the skills an AI needs in weapon systems.

3.2 CHARACTERISTICS MAKING CONTROL DIFFICULT AND NEW APPROACHES

If AI is indeed the core element of autonomous weapons systems and learning capability is further enhancing the capabilities of these systems without the need for adaptation of hardware, then it can be concluded that ultimately the focus of arms control should not be on autonomy, but on weapon-controlling MLpAI. But, as will now be shown, MLpAI leads to completely new problems for arms control.

3.2.1 INTERCHANGEABLE EXTERIORS THROUGH INCREASED HARDWARE COMPATIBILITY

As discussed in chapter 3.1, the processes of development of hardware and software can be considered separately. However, since the interaction of the two levels must be coordinated in practical applications, it is not possible to achieve general interchangeability of hardware. The software must be able to communicate with the hardware. One way to increase compatibility with a large variety of hardware components is to have consistent standards and automatic driver updates. For more complex hardware systems, so-called “middleware” is used in robotics. The middleware governs the heterogeneity of hardware and applications through an additional layer. It facilitates the integration of new technologies, the use of sensor data and the interchangeability of components. The integration of machine learning can further increase the compatibility of middleware by allowing it to dynamically adapt to the system (Bennaceur et al. 2013). If such methods are also used in the development of weapons systems, MLpAI could be used in different systems without complex adjustments. MLpAI could equally well control a drone, an underwater vehicle, a rocket, or other robotic weapons. Thus, the element of new weapons technologies cannot be associated exclusively with any specific weapons system.

A common approach of arms control is the quantitative limitation of a weapon’s carrier system. If MLpAI is understood as a weapon or its multiplier, conventionally armed drones, robots, etc. represent the carrier systems. However, limiting the carrier system would merely cause the transfer of the MLpAI to another system. Navigation, target identification, and action could take place on all systems by means of a similar setup, thus providing the compatibility required for transfer. Thus, as a destructive core element of weapons systems MLpAI cannot be identified through an externally visible system, which thus eliminates one of the usual approaches to arms control.

3.2.2 EXCHANGEABILITY OF EXTERNALLY VISIBLE CAPABILITIES THROUGH SOFTWARE UPDATES AND OPEN SOFTWARE ARCHITECTURE

A typical feature of computer programs are updates, which are intended to close security holes, add features, or modify components. With the increasing use of software-based weapons systems, updates are also necessary here. Highly engineered fighter aircraft exemplify this: The *US Air Force* updated the software of the F-22 fighter aircraft in order to make it capable of firing newer weapons, identifying targets better, and thus performing a wider range of assault missions (Osborn 2017). In addition, weapon systems that use MLpAI can be equipped with additional functions without changing the hardware. To further increase the flexibility of this process, open software architecture can be introduced. Such architecture can be found in smartphones: So-called “apps” are applications that allow the system to add, remove or update components without changing the main program. In the armaments industry, this system is already being used in the F-35 fighter jet. The Israeli military imports this aircraft and, using open architecture, can adapt it to its own requirements without changing the central software (Adams 2016). Other US defense companies are also developing open software architecture for their own products. If a group-wide standard were established, applications could be used flexibly, regardless of the type and design of the weapons system (Hagen et al. 2012: 6).

One approach to arms control is to limit the capabilities of weapons. The *Nuclear Test Ban Treaty*, once in force, prohibits nuclear explosions for civilian or military purposes. In this case, arms control relies on the explosive power of nuclear weapons. However, MLpAI functions can be flexibly added to or removed from weapons systems. Identification and restriction of weapons systems are not possible on the basis of visible capabilities. Even during inspections, critical functions could be added or removed for a short time.

3.2.3 INTRANSPARENT INTERNAL MODE OF OPERATION THROUGH COMPLEX REVERSE ENGINEERING AND LACK OF EXPLAINABILITY OF THE LEARNING MODEL

If an agreement on the functioning of a computer program is to be verified, the source code of the program can be analyzed. If only the completed system is available, for example, an autonomous drone, the source code cannot be easily extracted. Programs are typically written in a so-called higher-level language, which is usually converted into machine language by a compiler.⁷ This transformation destroys meta-information, making it difficult to reverse the process. Special programs can be used to reconstruct the source code through so called reverseengineering (Eilam 2005). Nevertheless, an initial hurdle to control and verification of digitally controlled weapons systems is created solely by the basic architecture of software. This already applies to today's weapons systems, even without machine learning.

In machine-learning applications, however, a second level of intransparency is added. As described in chapter 2.2, inherent characteristics of the different learning methods determine its transparency. Until now, scarcely-explored methods would need to be added to the weapons system to justify its capabilities. Apart from after-the-fact clarification, prior determination of capabilities is impossible as long as the model learned is not visible.

Another common approach for arms control is the mode of operation of, for example, nuclear weapons, anti-personnel mines and cluster munitions. In the case of nuclear weapons, the use of nuclear energy for an explosion is controlled. Anti-personnel landmines and cluster munitions are prohibited, as their mode of operation cannot distinguish between combatants and civilians. In the case of MLpAI, the mode of operation is made intransparent by the two levels already described. This lack of transparency eliminates a further approach of arms control, since MLpAI in weapons systems cannot be defined and restricted on the basis of the mode of operation.

3.2.4 TECHNICAL APPROACHES THAT ENABLE ARMS CONTROL

If – as concluded in previous chapters – the existence, hardware, mode of operation, and capabilities of MLpAI do not provide a reliable approach that allows arms control to be carried out – then this

⁷ Higher-level language can also be interpreted in real time. Direct interpretation makes testing and modifying of the code easier. Nonetheless, compiling is still customary, as it increases the efficiency of the code.

must be achieved by implementing control measures on arms development or deployment. Development as an approach is addressed by the concept of preventive arms control. In this approach, militarily applicable technologies, substances or systems are identified and banned or regulated during the development or testing phase (Altmann 2008):

“In concrete terms, preventive arms control aims at limiting, suspending, or terminating related research and development processes and/or prohibiting military options based on their implementation in weapons (systems).” (Neuneck/Mutz 2000: 109)

In this context, a register for military research and development, which would detect armament risks at an early stage, is conceivable (Müller 2000). But this approach requires a high degree of transparency in the development process, which could then be copied by other parties far more easily than current technologies. In addition, the overlap with civilian AI research makes it difficult to clearly clarify intentions (Bostrom 2017).

The application of AI in weapons systems could be achieved by restricting the strategic and tactical goals of the system and the associated options for use (Kahl/Mölling 2005: 350). Such mission objectives and the actions of the system could be recorded in a kind of black box. A “glass box”, as proposed by Gubrud and Altmann, could increase transparency immensely:

“A time slice of the data stream immediately prior to and including the selection and engagement commands could be designated as the primary record of the engagement. This record would be held by the state party, but a cryptographic code called a ‘hash’ of the record would be recorded by a ‘glass box’ [...] together with a time stamp of the moment the engagement command was issued. The hash would serve as a digital seal of the engagement record; if even a single bit of the record were later altered, the hash would not match.” (Gubrud/Altmann 2013: 6)

In the event of suspicion of illegal acts of war, the state would have to hand over the recorded data from the box to an international verification authority.

3.3 CONCEPTUAL REFLECTION: STRATEGIC INSTABILITY AND PROLIFERATION

After identifying the relevance and hurdles associated with MLpAI as an object of control, this chapter will apply arms control theory to working out the consequences of the use of MLpAI. The goal of stabilizing inter-state relations is to be achieved by preventing military escalation and an arms race. But humans’ options for achieving de-escalation by means of arms control would vanish with the autonomy of weapons systems. In addition, MLpAI could contribute to vertical proliferation – further military-technological development and improvement of existing capacities – and thus to a new arms race. In addition, MLpAI development can be used equally well for civilian and military

purposes, and thus contribute to horizontal proliferation – dissemination of military-technological knowledge and weapons systems among state and non-state actors.

3.3.1 CRISIS INSTABILITY: LACK OF DE-ESCALATING HUMAN NATURE

Using MLpAI with sophisticated perceptual, learning, and inference skills in weapons systems can result in a high degree of autonomy. This endangers the overall armaments policy goal of crisis stability, since the moderating factor constituted by human beings is minimized. The essence of this factor is that human beings slow things down:

“Despite modern communications and electronic data processing, officials are still limited by ordinary human intelligence, the conventional speed of spoken language, reading motions of the eye and the emotional accompaniments of responsibility in a crisis.”
(Schelling/Halperin 1961: 27)

During this brief time, humans have three de-escalation options:

- Validation of the machine’s report or recommendation
- Communication with the opponent to seek negotiations or explanations
- Weighing up moral and legal implications

Arms control relies on these options by delaying the operational readiness of weapons systems. “Many weapons limitations seem to be oriented, implicitly if not explicitly, towards the pace of decision” (Schelling/Halperin 1961: 27). In conventional arms control this is achieved by making it illegal for missiles to be permanently ready to launch, requiring warheads to be stored separately from missiles, or requiring submarines equipped with missiles to remain in coastal waters. In several cases during the Cold War, nuclear escalation was also prevented by people identifying a technical false alarm (Schlosser 2013). Arms control not only exploits this human factor but strengthens it through trust-building measures. These measures aim at confidence-building among people by, among other things, exchanging various types of information, allowing the presence of foreign observers at military exercises, maintain exchange programs for officers and trainees, or hotlines for crisis situations (Goldblat 2002: 11).

If the autonomy of a weapons system is at a level that does not permit human monitoring and intervention, the automatic escalation of a situation becomes more likely. An example of an incident caused by AI is the *flash crash* on the New York Stock Exchange in May 2010, when market manipulation set off a downward spiral of sales by computer-controlled high-volume traders. As a result, officials set up safety measures (CFTC/SEC 2010). Autonomous weapons systems could become involved in similar situations, where an error or chance causes exceptionally rapid escalation. A fallback mechanism to a human decision maker in the event of unexpected behavior by the AI would be a stabilizing precaution that would prevent a violent escalation (Scharre 2016: 38–39). Such a mechanism is also necessary so that confidence-building measures do not become redundant:

Trust that has been built up is less valuable if the weapons systems make their own decisions and human judgments of the opposing party no longer exert any influence.

But MLpAI in particular needs human monitoring as long as systematic learning errors and susceptibility to manipulation have not been eliminated (see Chapter 2.2). For example, if MLpAI has to distinguish between combatants and civilians, a faulty or manipulated learning model can lead to false classification. Enemy combatants could protect themselves by altering optical features on their clothing or weapons so that they are misclassified. MLpAI, the core of autonomous weapons systems, lacks the de-escalating human nature which could prevent such a scenario. The situation can be summed up in the words of Altmann and Sauer:

“Speed is undoubtedly a tactical advantage on the battlefield, and humans are slower than machines. But strategic stability is essential for survival. When it comes under threat, some remainder of human slowness is a good thing.” (Altmann/Sauer 2017: 136)

3.3.2 ARMS RACE INSTABILITY AND VERTICAL PROLIFERATION: THE RISK OF A REVIVED EXTERNALLY-LED ARMAMENT DYNAMIC

An interaction between the development of MLpAI and the armaments dynamics of a country can arise that can be guided both internally and externally. From the perspective of internally directed armaments dynamics, the source of this interaction is power relationships within a society. AI research is part of a major technical phase in military development, most clearly visible in the US (Neuneck/Alwardt 2008). Especially in democracies, rearmament is being promoted in order to develop the best weapons systems possible in order to minimize the risk of losses in warfare (Shaw, 2005: 79; Schörnig 2008). In addition, these states also have strong military-industrial actors, who use their domestic political influence to promote ongoing military development (Müller/Schörnig 2006: 106).

The perspective of externally directed armament dynamics sees their origin as lying in relationships among two or more states. In one form, states are seen as striving for intensive military technological development in order to gain technological superiority over other states. On the basis of the Cold War nuclear arms race, Matthew Evangelista (1988) shows that the development of technological innovations in weapon systems took a very different course in the US and the USSR. While in the US innovations sprang from the strong civil society, in the USSR innovation was centralized and reactively specified by the state. Evangelista’s model is helpful in understanding current innovations in AI research: In keeping with the bottom-up approach, US research is largely funded by corporations in the private sector (Bughin et al 2017: 10). In Russia and China, on the other hand, governments are promoting development. As Vladimir Putin says:

“Artificial intelligence is the future, not only for Russia but for all humankind, [...]. Whoever becomes the leader in this sphere will become the ruler of the world.” (Lant 2017)

In addition, the *Russian Military Industrial Committee* announced that by 2025 it would replace 30 percent of military technology with robotics and autonomous systems (Association of the United States Army 2017: 1). China also released plans promising US\$150 billion over the next few years, in order to make China an innovation center for AI by 2030. As Chinese president Xi Jinping says:

“We need to speed up building China into a strong country with advanced manufacturing, pushing for deep integration between the real economy and advanced technologies including the internet, big data, and artificial intelligence.” (Yu/Jing 2017)

The extraordinary progress of Chinese AI research is to be further expanded and the military is seeking civilian cooperation to integrate AI into the armed forces (Kania 2017). The investment programs of Russia and China could reduce the US military and technological advantage.

If MLpAI-controlled weapons systems deliver a high strategic advantage comparable to nuclear weapons – even if of a different kind – and the government investments in military applications that have been announced are realized, a new arms race among at least the three actors China, Russia, and the USA could be initiated.

3.3.3 DUAL-USE AND HORIZONTAL PROLIFERATION: UNCONTROLLABLE DISTRIBUTION AND USE

Another aim of arms control is to curb the proliferation of weapons (systems) and military-technical knowledge. This objective is particularly threatened by the *dual use*⁸ problem, because advances in ongoing civilian development can be used for military purposes:

“[O]pen-sourcing the code for autonomous weapons seems undesirable, and we have not heard anybody calling for that to be done. But basic research in AI is typically not application-specific in this way. Rather, to the extent that it succeeds, it will deliver algorithms and techniques that could be used in a very wide range of applications.” (Bostrom 2017: 137)

Even if the source codes of future (semi-) autonomous weapons systems are not publicly available, public basic research⁹ is not application-specific and could be exploited for illegal purposes. A variety of MLpAI applications and programming frameworks are freely available.¹⁰ These open source projects are used on the one hand to develop the defense industry and on the other hand in the development of civil projects. The more diverse the fields of application of a single piece of civilian MLpAI, the more likely is its use in weapons systems. In the case of nuclear weapons technology, the dissemination of knowledge was limited by the test stop norm, as the development of an advanced

8 “The trade in dual-use items – goods, software and technology that can be used for both civilian and military applications and/or can contribute to the proliferation of Weapons of Mass Destruction (WMD) – is subject to controls to prevent the risks that these items may pose for international security.” (European Commission 2017)

9 A large number of well-known researchers publish their latest results (without source code) at <https://arXiv.org>.

10 A selection of freely accessible learning architectures: Apache Singa, H2O, TensorFlow, Torch, Accord.NET.

nuclear weapon requires several test runs. The tests must be performed on a scale that inevitably emits observable seismic waves, hydroacoustic signals or radionuclides. In the case of MLpAI, banning tests would not be effective, as small-scale functional tests or simulations – which do not emit detectable signals – can be performed and then scaled to many systems. The evolution of basic MLpAI into a weapon-ready application is a relatively small hurdle and makes arms control to prevent the spread of high-risk technology a distant goal.

4. MACHINE LEARNING-POWERED ARTIFICIAL INTELLIGENCE IN VERIFICATION MEASURES

“Trust, but verify” is an oft-quoted dictum in arms control, for an agreed-upon limitation of armaments does not liberate a state from the fundamental mistrust of other states. Only the verification – that is, the examination of whether states are complying with an arms control treaty – can reduce mistrust and strengthen the aspiration of every state to security.

The three objectives of verification are, *firstly*, to create transparency and thus to recognize treaty breaches early, in order to initiate diplomatic, military or economic measures. Through the prospect of these reactions, verification measures should discourage breaches of treaties. In addition to the deterrent function, these measures should *secondly* also build trust between parties. Confirming that all member states are respecting a treaty builds confidence in the value of arms control for the protection of national interests (Goldblat 2002: 309). *Thirdly* verification measures allow falsely accused states to demonstrate their compliance with a treaty. When such an allegation is formulated, evidence from verification and intelligence sources must be collected and evaluated. If the allegation proves to be true, Member States may adopt measures within the regime or refer the breach to the United Nations Security Council (Müller/Schörning 2006: 150–153). Failure of diplomatic measures to correct breaches of a treaty may result in embargoes, sanctions or the threat of military force.

The potential measures make it clear that the validity and quality of the evidence collected is essential for the assessment. Off-site inspections (satellites, airplanes, radar and other sensor systems) and on-site inspections make it possible to monitor the armed forces, weapons or activities of the Member States. The inspections can be “conducted either in a systematic manner – continuously or periodically – or ad hoc, as decided by the verifying body, or upon challenge, as a result of a specific demand” (Goldblat 2002: 310). The verification process involves collecting, processing, and analyzing data to derive actionable information. MLpAI can be used to increase validity and quality or improve the efficiency of human analysts in one or all three of these steps.

4.1 APPLICATIONS AND TECHNICAL POTENTIAL

The potential of MLpAI to collect, process, and analyze data for verification is evident from applications or research already today. Various data sources are suitable as starting points for analyzing these cases: Satellite imagery from space, inspections on the ground, or global networks of sensors.

4.1.1 FROM SPACE: AUTOMATED REMOTE SENSING WITH SATELLITE IMAGES

For example, the analysis of satellite imagery plays an important role in arms control regimes for non-proliferation of nuclear weapons and energy, or for the limitation of short- and medium-range missiles. The information obtained from satellite imagery – the process is also known as “remote sensing” – is used to verify treaty terms (Patton et al 2016, Niemeyer/Ruthowski 2016). To increase the validity and conclusiveness of these analyses, MLpAI can be used to identify objects and changes over time on the images.

Intensive use is made of aerial photographs by, for example, the *International Atomic Energy Organization* (IAEO), which monitors maintenance of the *Nuclear Non-Proliferation Treaty*. For monitoring nuclear facilities, the IAEA uses the spatial, temporal and multispectral dimensions provided by satellite imagery. The aim of the analyses is to detect undeclared facilities for production of highly enriched uranium or plutonium, to track the use and processing of heavy metals, and to identify objects of interest for on-site inspections. For this, indicators such as temperature radiation or optical changes to buildings are used to monitor the expansion or the intensity of use of a facility (Truong et al. 2005; Johnson et al. 2014).

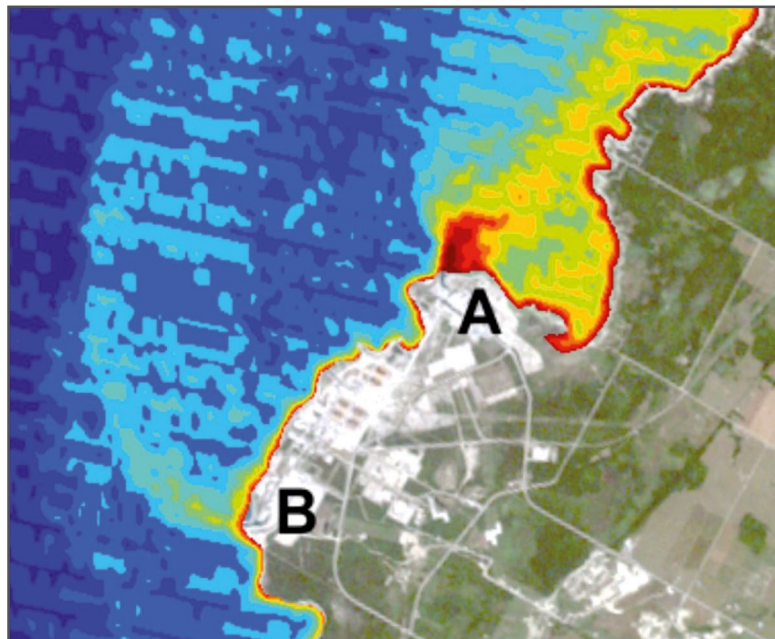


Fig. 2: A thermographic image of the Bruce nuclear power plant in Canada shows that unlike Block B, Block A is in operation. (Truong et al. 2005: 3)

One concrete challenge, for example, lies in the identification and monitoring of uranium mills, which produce the so-called *yellow cake* in the nuclear fuel cycle of natural uranium, a precursor of uranium used in nuclear power plants or nuclear weapons. Since these look very similar to completely harmless copper mills on satellite imagery, an AI developed by researchers uses characteristics of individual buildings and the size of the complex to classify the installations (Sundaresan et al. 2017). The AI carries out the classification using a human predefined decision tree. Although the IAEA has always adapted its actions to technological developments in the past, there is currently no evidence that it uses machine learning in the practical analysis of aerial photography.¹¹ It will, however, continue to be listed as undeveloped innovation in the updated 2018 research plan, but without high priority (International Atomic Energy Agency 2018: 15).

Commercial suppliers already offer usable applications: The geographic information system ENVI operates a module that allows researchers and analysts to load a sample image of an object that is being searched for, such as tanks for chemicals, as training data, and then have the program search for aerial photographs. In addition to application in urban planning, nature conservation and forestry, the manufacturer explicitly advertises the finding or locating of military vehicles, landing zones or buildings (Harris Cooperation 2017). Other suppliers¹² have adopted the business model of AI-based evaluation of commercial satellite imagery. *"We could not have done this five years ago,"* says CEO Pavel Machalek of SpaceKnow (Dillow 2016). And the potential is still great, because advances in the combination of computing power, machine learning and satellite images are only just beginning.

Due to US restrictions, only relatively coarse-grained satellite images with a resolution of 30 to 40 cm per pixel are available to non-state actors (Shalal 2014), whereas US spy satellites have a presumed resolution of 15 cm (Krebs 2017). Nevertheless, there are tremendous opportunities for MLpAI to use these images, because MLpAI is able to compensate any drawbacks resulting from poorer image resolution. Apart from small arms and light weapons, commercial machine learning can also identify large pieces of military equipment on low-resolution commercial satellite images.

Depending on the size of the object, MLpAI can also recognize unmistakable features and track an object across multiple images. Conspicuous changes to nuclear, chemical or biological factories can be analyzed automatically and anomalies communicated to the analysts of control regimes. This outstanding anomaly and pattern recognition feature can make the use of MLpAI in satellite-based verification more valid, thus strengthening it as an instrument.

11 Within the newly developed IAEA Geospatial Exploitation System, analysis tools from commercial providers should also be available (Balter 2014: 6). Machine learning could possibly be used there.

12 Satellite operator DigitalGlobe offers a programming framework for finding objects at: <http://deepcore.io/>. The SpaceKnow (<https://spaceknow.com/>), Orbital Insight (<https://orbitalinsight.com/>) and Descartes Labs (<https://www.descarteslabs.com/>) start-ups offer online platforms for object and pattern recognition. A simplified function can be tried out on the sub-page <https://search.descarteslabs.com/>.

4.1.2 ON THE GROUND: ARMS TRADE AND CIVIL CLEARANCE OF UNEXPLODED EXPLOSIVE DEVICES

The trade in conventional weapons and thus among other things compliance with the *Arms Trade Treaty* could be monitored using MLpAI. The treaty obliges states to track arms exports to their destinations and to prevent potential transfers to states that commit human rights violations or violations of international humanitarian law. In order to be able to guarantee this for transit goods too, the countless freight containers that are loaded, unloaded and reloaded daily in harbors must be checked (Holtom/Bromley 2011). Due to the huge volume of trade, human monitoring can at best involve sampling based on known risk factors such as origin, destination and declared content. But an MLpAI would be able to use X-rays to identify armaments within countless sealed containers, as a research team from University College London has already demonstrated (Jaccard et al. 2016). While humans would be able to interpret the images of contents created by the sensors, machine learning can capture the data much more quickly, recognize patterns in the data, and output a probability estimate of the type of content. In the training phase of the learning model, the MLpAI would have to be confronted with numerous different scenarios as well as the correct answer – e.g., “warhead” or “no warhead.”

A second application example comes from humanitarian arms control: Anti-personnel mines and cluster munitions remain in the ground even after the conflict has ended, endangering the civilian population. The humanitarian arms control treaties, the *Ottawa Convention* and the *Convention of Cluster Munitions*, impose both a general prohibition and an obligation to remove weapons already in place on the parties to a conflict. This elaborate process requires trained teams which systematically pinpoint the areas involved, detect undetonated mines with metal detectors, and then detonate them in controlled explosions. Highly efficient methods have been and are being developed to meet the great challenge of finding the objects. A highly promising new approach combines ground radar with machine learning. As can be seen in the figure, radar images of an anti-personnel mine and of a flat beverage can are not easy to distinguish visually. To this end, various development teams trained an MLpAI using such input data (Núñez-Nieto et al. 2014; Seiffert et al. 2013). This method makes it possible to mount the system on an off-road vehicle and scan the ground a few meters in front of the car. The aim is to reduce the rate of accidents or false alarms compared with other tracking methods.

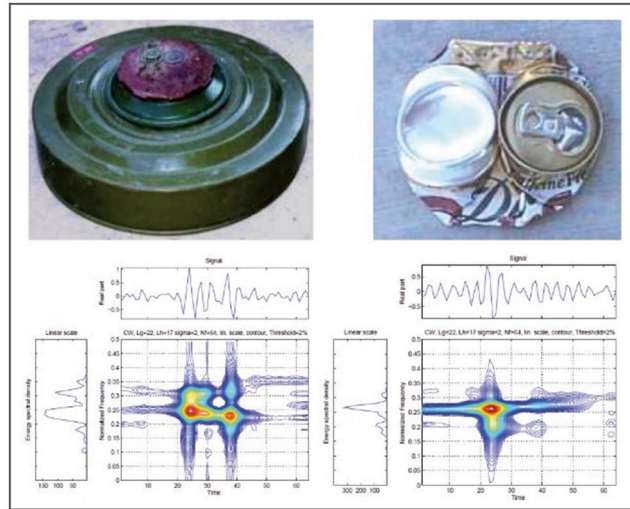


Fig. 3: Comparison of sensor data for an anti-personnel mine and a beverage can (Sun/Li 2005: 3)

The two cases in this chapter have in common that the MLpAI receives immediate sensor data from the environment for identifying patterns. In the detection of armaments or anti-personnel landmines, the interest lies in the distinction from other objects based on incomplete or unstructured data. Since these cases are actively being tested, implementation and application is likely to show that MLpAI is more efficient than legacy systems. If these developments are generalized and extrapolated, there is a considerable potential for material transparency: Activities or objects that violate arms control agreements can only be partially shielded against discovery, even with the utmost care. The small amount of unwanted information that penetrates the shielding can be interpreted by humans or statistical models either not at all or too slowly, but MLpAI can achieve a new level of transparency at this point.

4.1.3 A WIDE RANGE OF SENSORS ENABLES “MEASUREMENT AND SIGNATURE INTELLIGENCE”

The processing of electro-optical, electromagnetic, acoustic and geophysical data as well as nuclear, biological and chemical trace elements is called *measurement and signature intelligence* (MASINT) in intelligence work (Aid 2014: 120–122). The basic idea of MASINT is that the object being investigated consistently identifies a unique signature across different information carriers (radiation, sound, substances, etc.). The analysis of this information can be used for verification measures.

A control method that uses acoustic and seismic sensor data was developed by the Bochum Verification Project. The measuring instruments developed in the project were placed at intervals of up to 200m to provide a line of sensors that also detected vehicles off the roads. The measuring devices were able to classify passing military vehicles into different vehicle categories based on acoustic and

seismic signals (Hochmuth 2003). The more promising acoustic recognition was able to differentiate between five chain and wheeled vehicles at a time on the basis of engine noise (Altmann et al. 2002). The varying success rate of 69% to 98% could be stabilized at a high level through MLpAI and the resilience to disruptive noise, such as a rain shower, improved. For example, a line of such meters could control the quantitative characteristics of a particular vehicle type. By contrast, the identification of individual vehicles may prove difficult. Although the engine noise varies slightly depending on the vehicle, this can change due to wear or repairs, making recognition impossible.

In another promising area of application, MLpAI may well be used to verify compliance with the *Comprehensive Nuclear-Test-Ban Treaty* (CTBT). The existing *International Monitoring System* (IMS) exemplifies conventional processing of an extremely high volume of sensor data. By capturing the data in a worldwide network of measurement stations, the IMS can diagnose nuclear explosions. Seismographs, hydroacoustic sensors, infrasonic sound stations and radionuclide detectors generate large amounts of data. Central data processing in Vienna analyzes and reduces the raw sensor data in order to distinguish signals from the constant noise and to classify them as either “nuclear explosion” or “not nuclear explosion” (Russell et al. 2010: 32). This process cannot yet be performed completely automatically. The analysts still have to laboriously rework the results of the automatic systems, since the automatic processing is prone to errors due to heavy noise, incorrect classification or false associations. In order to further optimize the results, new methods were already proposed by various project groups in 2009 at the *International Scientific Studies* (ISS) conference in Vienna. The project groups agreed unanimously on the problem: The data recorded by the sensors of the IMS is too complex for the performance of a conventional statistical analysis based on linear discrimination – a linear function that defines the boundary between groups. All four project groups are testing the ability of MLpAI to make non-linear discriminations. The project groups have already used classified data from the last ten years as “ground truth” (training set for the machine learning algorithm). The prototypes analyzed seismic data (Kleiner et al. 2009), hydroacoustic data (Tuma/Igel 2009), infrasound data (Procopio et al. 2009), and radionuclide data (Stocki et al. 2010). All prototypes were able to make the analysts’ task easier and make more accurate automated evaluations. Nevertheless, the researchers came to the conclusion that the systems are not yet operational and would need to be further refined. Overall, these research projects and improved further development (Arora et al. 2013) have demonstrated the high potential of MLpAI for verification in arms control based on the IMS data.

Analysis of large volumes of physical data – translated by sensors – can be greatly improved by MLpAI compared with human analysts. In particular, verification of the nuclear test ban will benefit tremendously from MLpAI in the short term, since the data is already structured by the sensors and is thus easier to analyze. MLpAI makes this type of verification much more attractive and could result in other weapons being monitored in a similar way.

4.2 CONCEPTUAL REFLECTION: IMPROVEMENTS IN EXISTING METHODS

The cases in the previous chapter show that MLpAI has the ability to strengthen arms control, especially its verification measures. Since the end of the East-West conflict, the significance of verification measures has stagnated due to new types of weapons technology and isolated breaches of the treaty (Pilat 2002: 85–87). MLpAI would have the potential to make these measures more relevant once again, provided that states muster the appropriate political will to make an effort here.

The further technical development of verification measures, if implemented, would allow greater transparency in the monitoring of control items. The realization of improved verification would also strengthen the goal of deterring a secret breach of the treaty. The two other objectives of verification measures – building trust and demonstrating compliance with the treaty – could also benefit from increased transparency resulting from the use of MLpAI.

As shown in the previous chapters, the use of MLpAI has already been tested in feasibility studies and shown to be effective in verification measures. Nevertheless, machine learning algorithms are still facing major challenges. Depending on the form of the input data, they may need to “perceive” and interpret visual and other sensory impressions. They have to recognize patterns in digital data and complete an analysis using existing knowledge. Perception, pattern recognition, and generation of knowledge can be improved through machine learning because human constraints on the analysis program cannot achieve the necessary complexity. The adaptability of machine learning allows programs to develop in parallel with the control objects when characteristics of the objects being analyzed change over time. MLpAI is able to distill knowledge from the highly complex and changing environment of verification actions, enhancing arms-related transparency. In addition, MLpAI allows greater scalability – a significant increase in the number of analyses performed. As Robert Cardillo, Director of the National Geospatial Intelligence Agency, states:

“If we were to attempt to manually exploit the commercial satellite imagery we expect to have over the next 20 years, we would need eight million imagery analysts.”
(Cardillo 2017)

At the same time, MLpAI offers other advantages over human analysts:

“Technology does have advantages over human inspectors. It can operate continuously and at a constant level of observation. Its data is not comparable. It can be limited to detecting treaty-relevant information, while ignoring other types of information.”
(UNIDIR/VERTIC 2003: 27)

But whether MLpAI is to be used as an official verification tool in traditional arms control is a political decision member states must make. For political reasons, technical verification possibilities have already been artificially restricted:

“The INF Treaty [...] permitted an x-ray to be taken of missile canisters to determine the type of missile inside, but the machinery was set to a certain resolution so that sensitive design information could not be obtained.” (UNIDIR/VERTIC 2003: 27)

The technical measures in the *Open Skies Treaty* – a treaty aimed not at restricting weapons but at building trust – were deliberately limited to a low standard: During the overflights permitted by the treaty, aerial photographs with a maximum resolution of 30 centimeters are permitted, although commercial satellite images already provide the same or better resolution. In addition, in the process of modernization from analogue to digital cameras, care was taken not to exceed the maximum resolution specified in the treaty (Britting/Spitzer 2005). Technological advances in traditional arms control treaties are not in the best interest of those nations that already have a head start in intelligence gathering. The use of MLpAI can – as has already happened with other traditional verification measures – be prevented or limited by political will or in the interest of military secrecy.

In humanitarian arms control, by contrast, development-inhibiting interests of nation states are unlikely to be expected. Pilot projects for the detection of anti-personnel mines and cluster munitions (see Chapter 4.1.2) show the immense benefit of rapid clearance of such non-discriminatory weapons. This principle could also be used by actors monitoring the international trade in small arms. For example, the *itrace* project tracks the distribution of such weapons through field investigations. Investigators could be supported by MLpAI, given that it can identify behavior patterns in data or unmistakable visual features, and thus link individual weapons to specific weapon inventories.

5. MACHINE LEARNING-POWERED ARTIFICIAL INTELLIGENCE IS PART OF THE PROBLEM AND PART OF THE SOLUTION

In this report, it became clear that in conventional weapons control, control of weapons of mass destruction and of modern weapons systems, MLpAI can be both part of the problem and part of the solution. As a control measure, by its very nature MLpAI eludes many approaches to imposing qualitative or quantitative limitations. Thus, it increases the relevance of alternative methods aimed at achieving overall military transparency or at confidence-building measures. It is precisely in these methods that MLpAI can again be used as a verifying instrument. Accurate and comprehensive information processing allows MLpAI to create greater transparency, verify compliance with treaties, and reinforce trust between parties.

From the perspective of arms control theory, the use of MLpAI in weapon systems and verification measures results in fragile strategic stability. Increased use of MLpAI in weapons systems can jeopardize strategic stability by minimizing the de-escalating human nature and promoting a technological arms race. The uncontrolled proliferation of basic MLpAI lowers the barriers of using MLpAI in weapons systems. Both these theoretical considerations and the realization that MLpAI will be the core element of future autonomous weapons systems are forcing limitations in technology by means of arms control. But qualitative or quantitative control on the basis of external appearance, externally recognizable functions, or internal operating principles is not possible in the case of MLpAI.

Traditional approaches reach their limits at this point and the only remaining possibilities are to monitor and restrict the MLpAI during development or deployment. If control of deployment through a “glass box” or preventive control of the development process is not enforceable, qualitative solutions could be sought at a higher level. Overall military capabilities and strategies could be presented more transparently and confidence-building measures intensified (Schörnig 2015).

Using MLpAI in verification measures can improve strategic stability, as it is to be expected that MLpAI will enable technical monitoring tools to provide information processing that is far more accurate and comprehensive. However, this opportunity is not without obstacles: In the current state of development of machine learning, transparency and protection against external manipulation need to be improved, in order for it to be regarded as a valid verification method. These technical requirements must be met in order to build trust in the method and thus also between states. The potential of MLpAI as a verification method can be seen in numerous prototypes and early applications. These show that the ability of machine learning can be used in the analysis of optical, thermal and topographical satellite imagery, sensor data in on-site inspections, civilian arms flights, and unstructured data from a network of measuring stations.

Whether the two phenomena analyzed here will have a significant impact on arms control depends on the environment in which the AI is to operate. Dividing the environment into types (Russell/Norvig 2010: 46) shows that real-world actions must take into account far more unpredictable environmental factors than are present in a digital environment. AI research still needs more years before it will be able to make it possible for weapons systems to take autonomous action in the real world. However, machine learning can already be used for verification, because the data to be analyzed is already available in digital form or can be interpreted according to a predetermined pattern, without having to be trained to deal with sudden changes in the environment. Timely deployment is critical, as MLpAI can already raise arms control to new levels of capacity before it is forced to face the new challenge of MLpAI-controlled weapons systems.

- Adams, Eric 2016: Why Only Israel Can Customize America's F-35 (at Least for Now), <https://www.wired.com/2016/05/israel-can-customize-americas-f-35-least-now/>; 04.12.2017.
- Aid, Matthew M. 2014: Measurement and Signature Intelligence, in: Dover, Robert/Goodman, Michael S./Hillebrand, Claudia (Eds.): *Routledge Companion to Intelligence Studies*, London, 114–122.
- Algorithm Watch 2017: Antworten auf den Fragenkatalog für das Fachgespräch zum Thema „Künstliche Intelligenz“ des Ausschusses Digitale Agenda vom 22.03.2017.
- Altmann, Jürgen 2008: Präventive Rüstungskontrolle, in: Becker, Una/Müller, Harald (Eds.): *Rüstungskontrolle im 21. Jahrhundert*, Berlin, 105–125.
- Altmann, Jürgen/Linev, Sergey/Weiß, Axel 2002: Acoustic–seismic Detection and Classification of Military Vehicles – Developing Tools for Disarmament and Peace-keeping, in: *Applied Acoustics* 63: 10, 1085–1107.
- Altmann, Jürgen/Sauer, Frank 2017: Autonomous Weapon Systems and Strategic Stability, in: *Survival* 59: 5, 117–142.
- Arbatov, Alexei 2015: An Unnoticed Crisis. The End of History for Nuclear Arms Control?, <https://carnegie.ru/2015/06/16/unnoticed-crisis-end-of-history-for-nuclear-arms-control-pub-60408>; 19.07.2019.
- Arora, N. S./Russell, S./Sudderth, E. 2013: NET-VISA. Network Processing Vertically Integrated Seismic Analysis, in: *Bulletin of the Seismological Society of America* 103: 2A, 709–729.
- Artificial General Intelligence Sentinel Initiative 2017: A Working List. Definitions of Artificial Intelligence and Human Intelligence.
- Association of the United States Army 2017: Integrating Army Robotics and Autonomous Systems to Fight and Win, <https://www.ausa.org/publications/integrating-army-robotics-and-autonomous-systems>; 19.07.2019.
- Balter, E. 2014: Digital Declarations: The Provision of Site Maps under INFCIRC/540 Article 2.a. (iii), <https://conferences.iaea.org/indico/event/47/contributions/8862/contribution.pdf>.
- Ben-Ari, Mordechai/Mondada, Francesco 2018: *Elements of Robotics*, Cham.
- Bennaceur, Amel/Issarny, Valérie/Sykes, Daniel/Howar, Falk/Isberner, Malte/Steffen, Bernhard/Johansson, Richard/Moschitti, Alessandro 2013: Machine Learning for Emergent Middleware, in: Moschitti, Alessandro/Plank, Barbara (Eds.): *Trustworthy Eternal Systems via Evolving Software, Data and Knowledge. Second International Workshop, EternalS 2012, Montpellier, France, August 28, 2012, Revised Selected Papers*, Berlin, Heidelberg, 16–29.
- Bojarski, Mariusz/Del Testa, Davide/Dworakowski, Daniel/Firner, Bernhard/Flepp, Beat/Goyal, Prashoon/Jackel, Lawrence D./Monfort, Mathew/Muller, Urs/Zhang, Jiakai/Zhang, Xin/Zhao, Jake/Zieba, Karol 2016: End to End Learning for Self-Driving Cars, <https://arxiv.org/pdf/1604.07316v1.pdf>; 19.07.2019.
- Boston Dynamics 2018: Boston Dynamics. Changing Your Idea of What Robots Can Do, <https://www.bostondynamics.com/robots>; 05.01.2018.
- Bostrom, Nick 2017: Strategic Implications of Openness in AI Development, in: *Global Policy* 8: 2, 135–148.

- Boulanin, Vincent/Verbruggen, Maaïke 2017: Mapping the Development of Autonomy in Weapon Systems, Stockholm, https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_0.pdf; 19.07.2019.
- Britting, Ernst/Spitzer, Hartwig 2005: Der Open-Skies-Vertrag: Stand und Perspektiven, in: Neuneck, Götz/Mölling, Christian (Eds.): Die Zukunft der Rüstungskontrolle, Baden-Baden, 308–323.
- Bughin, Jacques/Hazan, Eric/Ramaswamy, Sree/Chui, Michael/Allas, Tera/Dahlström, Peter/Henke, Nicolaus/Trench, Monica 2017: Artificial Intelligence. The Next Digital Frontier?, <http://www.odbms.org/2017/08/artificial-intelligence-the-next-digital-frontier-mckinsey-global-institute-study/>; 19.07.2019.
- Cardillo, Robert 2017: GEOINT 2017 Symposium (Remarks as prepared for Robert Cardillo), <https://www.nga.mil/MediaRoom/SpeechesRemarks/Pages/GEOINT-2017-Symposium.aspx>; 19.07.2019.
- CFTC/SEC 2010: Findings Regarding the Market Events of May 6, 2010, <https://www.sec.gov/news/studies/2010/marketevents-report.pdf>; 05.01.2018.
- Courtland, Rachel 2016: DARPA's Self-Driving Submarine Hunter Steers Like a Human, <https://spectrum.ieee.org/autoton/robotics/military-robots/darpa-actuv-self-driving-submarine-hunter-steers-like-a-human>; 05.01.2018.
- Croft, Stuart 1996: Strategies of arms control. A History and Typology, Manchester, UK.
- Daniels, Jeff 2017: Mini-nukes and Mosquito-like Robot Weapons Being Primed for Future Warfare, <https://www.cnn.com/2017/03/17/mini-nukes-and-inspect-bot-weapons-being-primed-for-future-warfare.html>; 05.01.2018.
- Danks, David 2014: Learning, in: Frankish, Keith/Ramsey, William (Eds.): The Cambridge Handbook of Artificial Intelligence, Cambridge, UK, 151–167.
- Dillow, Clay 2016: What Happens When You Combine Artificial Intelligence and Satellite Imagery, <http://fortune.com/2016/03/30/facebook-ai-satellite-imagery/>; 26.11.2017.
- Eilam, Eldad 2005: Reversing. Secrets of Reverse Engineering, Indianapolis, USA.
- Ernest, Nicholas/Carroll, David/Schumacher, Corey/Clark, Matthew/Cohen, Kelly/Lee, Gene 2016: Genetic Fuzzy based Artificial Intelligence for Unmanned Combat Aerial Vehicle Control in Simulated Air Combat Missions, in: Journal of Defense Management 6: 1, 1–7, <https://www.omicsonline.org/open-access/genetic-fuzzy-based-artificial-intelligence-for-unmanned-combat-aerialvehicle-control-in-simulated-air-combat-missions-2167-0374-1000144.pdf>; 31.01.2018.
- European Commission 2017: Dual-use Export Controls, http://ec.europa.eu/trade/import-and-export-rules/export-from-eu/dual-use-controls/index_en.htm; 19.07.2019.
- Evangelista, Matthew 1988: Innovation and the Arms Race. How the United States and the Soviet Union Develop New Military Technologies, Ithaca, NY.
- Farrelly, Colleen 2016: Machine Learning by Analogy, <https://www.slideshare.net/ColleenFarrelly/machine-learning-by-analogy-59094152>; 09.01.2018.
- Franklin, Stan 2014: History, Motivations, and Core Themes, in: Frankish, Keith/Ramsey, William (Eds.): The Cambridge Handbook of Artificial Intelligence, Cambridge, UK, 15–33.

- Future of Life Institute 2015: Autonomous Weapons. An Open Letter From AI & Robotics Researchers, <https://futureoflife.org/open-letter-autonomous-weapons/>; 14.01.2018.
- Goldblat, Jozef 2002: Arms Control. The New Guide to Negotiations and Agreements, London.
- Goodfellow, Ian 2016: Deep Learning, Cambridge, Massachusetts, London, England.
- Goodfellow, Ian J./Shlens, Jonathon/Szegedy, Christian 2015: Explaining and Harnessing Adversarial Examples, <https://arxiv.org/pdf/1412.6572.pdf>; 19.07.2019.
- Gubrud, Mark/Altmann, Jürgen 2013: Compliance Measures for an Autonomous Weapons Convention. ICRAC Working Paper #2, https://www.icrac.net/wp-content/uploads/2018/04/Gubrud-Altman-Compliance-Measures-AWC_ICRAC-WP2.pdf; 19.07.2019.
- Gunning, David 2016: Explainable Artificial Intelligence (XAI). Broad Agency Announcement, <https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>; 07.12.2017.
- Hagen, Christian/Sorenson, Jeff/Hurt, Steven/Wall, Dan 2012: Software: The Brains Behind U.S. Defense Systems, https://www.atkearney.com/documents/10192/247932/SoftwareThe_Brains_Behind_US_Defense_Systems.pdf/69129873-eccc-4ddc-b798-c198a8ff1026; 19.07.2019.
- Harris Cooperation 2017: ENVI Feature Extraction Module, http://www.harrisgeospatial.com/Portals/0/pdfs/HG_ENVI_FX_module_data-sheet_WEB.pdf; 26.11.2017.
- Hawkins, Jeff/Blakeslee, Sandra 2004: On Intelligence, 1st Edition, New York.
- Hochmuth, Olaf 2003: Bochumer Verifikationsprojekt – Sensorstation 2000, <https://www2.informatik.hu-berlin.de/~hochmuth/bvp/>; 13.12.2017.
- Holtom, Paul/Bromley, Mark 2011: Transit and Trans-shipment Controls in an Arms Trade Treaty. <https://www.sipri.org/sites/default/files/files/misc/SIPRIBP1107a.pdf>; 19.07.2019.
- Human Rights Watch 2012: Losing Humanity. The Case against Killer Robots, <https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots>; 14.01.2018.
- International Atomic Energy Agency 2018: Research and Development Plan. Enhancing Capabilities for Nuclear Verification, Wien, Österreich.
- Jaccard, Nicolas/Thomas W. Rogers/Edward J. Morton/Lewis D. Griffin 2016: Automated Detection of Smuggled High-risk Security Threats Using Deep Learning, <https://arxiv.org/pdf/1609.02805.pdf>; 19.07.2019.
- Johnson, Michael R./Paquette, Jean-Pierre/Elbez, Julien 2014: New and Emerging Trends In Satellite Imagery, <https://www.iaea.org/safeguards/symposium/2014/home/eproceedings/sg2014-papers/000042.pdf>; 26.11.2017.
- Kahl, Martin/Mölling, Christian 2005: Die „Revolution in Military Affairs“ und die Bedingungen und Möglichkeiten für Rüstungskontrolle, in: Neuneck, Götz/Mölling, Christian (Eds.): Die Zukunft der Rüstungskontrolle, Baden-Baden, 341–353.
- Kania, Elsa B. 2017: Quest for an AI Revolution in Warfare. The PLA's Trajectory from Informatized to „Intelligentized“ Warfare, <https://thestrategybridge.org/the-bridge/2017/6/8/-chinas-quest-for-an-ai-revolution-in-warfare>; 14.01.2018.
- Keller, John 2015: DARPA TRACE Program Using Advanced Algorithms, Embedded Computing for Radar Target Recognition, <http://www.militaryaerospace.com/articles/2015/07/hpec-radar-target-recognition.html>; 01.12.2017.

- Kleiner, Ariel/Mackey, Lester/Jordan, Michael I. 2009: Machine Learning for Improved Automated Seismic Event Extraction, https://www.ctbto.org/fileadmin/user_upload/ISS_2009/Poster/DM-02A%20%28US%29%20-%20Ariel_Kleiner%20etal.pdf; 19.07.2019.
- Knight, Will 2017: The Dark Secret at the Heart of AI. No One Really Knows How the Most Advanced Algorithms Do What They Do. That could be a problem., <https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/>; 11.01.2018.
- Krebs, Gunter 2017: KH-11/Kennen/Crystal, http://space.skyrocket.de/doc_sdat/kh-11.htm; 25.11.2017.
- Kurakin, Alexey/Goodfellow, Ian/Bengio, Samy 2017: Adversarial Examples in the Physical World, <https://arxiv.org/pdf/1607.02533.pdf>; 19.07.2019.
- Lant, Karla 2017: China, Russia and the US are in an Artificial Intelligence Arms Race, <https://futurism.com/china-russia-and-the-us-are-in-an-artificial-intelligence-arms-race/>; 16.11.2017.
- Legg, Shane/Hutter, Marcus 2007: A Collection of Definitions of Intelligence, <https://arxiv.org/pdf/0706.3639.pdf>; 19.07.2019.
- Li, Wei/Gauci, Melvin/Groß, Roderich 2016: Turing learning. A Metric-free Approach to Inferring Behavior and its Application to Swarms, in: *Swarm Intelligence* 10: 3, 211–243.
- McCloskey, Paul 2017: What's AI, and what's not, <https://gcn.com/Articles/2017/03/10/defining-AI.aspx>; 15.11.2017.
- Morisse, Tom 2017: The Next Challenges of AI Research, <https://en.fabernovel.com/insights/tech-en/the-next-challenges-of-ai-research>; 18.12.2017.
- Müller, Harald 2000: Früherkennung von Rüstungsrisiken in der Ära der „militärisch-technischen Revolution“. Ein Register für militärische Forschung und Entwicklung, Frankfurt.
- Müller, Harald/Schörnig, Niklas 2006: Rüstungsdynamik und Rüstungskontrolle. Eine exemplarische Einführung in die Internationalen Beziehungen, Baden-Baden.
- Müller, Vincent C./Bostrom, Nick 2016: Future Progress in Artificial Intelligence: A Survey of Expert Opinion, in: Müller, Vincent C. (Ed.): *Fundamental Issues of Artificial Intelligence*, Berlin, 553–571.
- Neuneck, Götz/Alwardt, Christian 2008: The Revolution in Military Affairs, its Driving Forces, Elements and Complexity, https://ifsh.de/pdf/publikationen/IFAR_Working_Paper_13.pdf?asset_id=5489; 19.07.2019.
- Neuneck, Götz/Mutz, Reinhard 2000: Vorbeugende Rüstungskontrolle. Ziele und Aufgaben unter besonderer Berücksichtigung verfahrensmäßiger und institutioneller Umsetzung im Rahmen internationaler Rüstungsregime, Baden-Baden.
- Niemeyer, Irmgard/Ruthowski, Joshua 2016: Satellite Imagery Processing for the Verification of Nuclear Non-Proliferation and Arms Control, Bonn.
- Núñez-Nieto, Xavier/Solla, Mercedes/Gómez-Pérez, Paula/Lorenzo, Henrique 2014: GPR Signal Characterization for Automated Landmine and UXO Detection Based on Machine Learning Techniques, in: *Remote Sensing* 6: 10, 9729–9748, <http://www.mdpi.com/2072-4292/6/10/9729/pdf>; 11.01.2018.
- O'Neil, Cathy 2016: *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*, 1st Edition, New York.

- Osborn, Kris 2017: Air Force Upgrades F-22 Sensors, Weapons Software, <https://defensesystems.com/articles/2017/03/14/f22.aspx>; 03.12.2017.
- Papernot, Nicolas/McDaniel, Patrick/Goodfellow, Ian/Jha, Somesh/Celik, Z. B./Swami, Ananthram 2017: Practical Black-Box Attacks against Machine Learning (Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security, Abu Dhabi, UAE), Abu Dhabi, UAE.
- Park, Dong H./Hendricks, Lisa A./Akata, Zeynep/Schiele, Bernt/Darrell, Trevor/Rohrbach, Marcus 2017: Attentive Explanations. Justifying Decisions and Pointing to the Evidence, <https://arxiv.org/abs/1612.04757>; 19.07.2019.
- Parkin, Simon 2015: Killer Robots: The Soldiers that Never Sleep, <http://www.bbc.com/future/story/20150715-killer-robots-the-soldiers-that-never-sleep>; 05.01.2018.
- Patton, Tamara/Lewis, Jeffrey/Hanham, Melissa/Dill, Catherine/Vaccaro, Lily 2016: Emerging Satellites for Non-Proliferation and Disarmament Verification, https://nonproliferation.org/vcdnp/wp-content/uploads/2016/06/160614_copernicus_project_report.pdf; 19.07.2019.
- Pilat, Joseph F. 2002: Verification and Transparency: Relics or Future Requirements?, in: Larsen, Jeffrey A. (Ed.): Arms Control. Cooperative Security in a Changing Environment, London, 79–96.
- Procopio, Michael J./Young, Christopher J./Gauthier, John A. 2009: Applying Machine Learning Methods to Improve Efficiency and Effectiveness of the IDC Automatic Event Detection System, https://www.ctbto.org/fileadmin/user_upload/ISS_2009/Poster/DM-07A__US_-_Michael_Procopio_et.al.pdf; 19.07.2019.
- Ribeiro, Marco T./Singh, Sameer/Guestrin, Carlos 2016: „Why Should I Trust You?“. Explaining the Predictions of Any Classifier, <https://arxiv.org/pdf/1602.04938.pdf>; 19.07.2019.
- Russell, Stuart J./Norvig, Peter 2010: Artificial Intelligence. A Modern Approach, 3rd Edition, Upper Saddle River, NJ.
- Russell, Stuart J./Vaidya, Sheila/Le Bras, Ronan 2010: Machine learning for Comprehensive Nuclear-Test-Ban Treaty monitoring, in: CTBTO Spectrum: 14, 32–35; 20.11.2017.
- Russia Today 2017: Kalashnikov Develops Fully Automated Neural Network-based Combat Module, <https://www.rt.com/news/395375-kalashnikov-automated-neural-network-gun/>; 05.01.2018.
- Sauer, Frank 2016: Stopping ‘Killer Robots’: Why Now Is the Time to Ban Autonomous Weapons Systems, https://www.armscontrol.org/ACT/2016_10/Features/Stopping-Killer-Robots-Why-Now-Is-the-Time-to-Ban-Autonomous-Weapons-Systems#note04; 13.01.2018.
- Scharre, Paul 2016: Autonomous Weapons and Operational Risk, https://www.files.ethz.ch/isn/196288/CNAS_Autonomous-weapons-operational-risk.pdf; 19.07.2019.
- Schelling, Thomas C./Halperin, Morton H. 1961: Strategy and Arms Control, New York.
- Schlosser, Eric 2013: Command and Control. Nuclear Weapons, the Damascus Accident, and the Illusion of Safety, New York, NY.
- Schmidt, Hans-Joachim 2017: Hoffnungsvoller Neustart der konventionellen Rüstungskontrolle?, <https://blog.prif.org/2017/07/10/hoffnungsvoller-neustart-der-konventionellen-ruestungskontrolle/>; 14.01.2018.

- Schörnig, Niklas 2008: Casualty Aversion in Democratic Security Provision. Procurement and the Defense Industrial Base., in: Evangelista, Matthew (Ed.): Democracy and Security. Preferences, Norms and Policy-making, London, 14–35.
- Schörnig, Niklas 2015: From Quantitative to Qualitative Arms Control: The Challenges of Modern Weapons Development, in: Development and Peace Foundation/Käte Hamburger Kolleg/Centre for Global Cooperation Research (Eds.): Global Trends 2015. Prospects for World Society, 87–100.
- Searle, John R. 1980: Minds, Brains, and Programs, in: Behavioral and Brain Sciences 3: 3, 417–457.
- Seiffert, Udo/Abeynayake, Canicious/Jain, Lakhmi C./Tran, Minh D.-J. 2013: Detection of Targets in Characteristic GPR Sensor Data Using Machine Learning Techniques, https://www.techfak.uni-bielefeld.de/~fshleif/mlr/mlr_01_2013.pdf; 19.07.2019.
- Shalal, Andrea 2014: DigitalGlobe Gains U.S. Govt License to Sell Sharper Satellite Imagery, <https://www.reuters.com/article/digitalglobe-imagery/digitalglobe-gains-u-s-govt-license-to-sell-sharper-satellite-imagery-idUSL2N0OR2UX20140611>; 25.11.2017.
- Shalev-Shwartz, Shai/Ben-David, Shai 2014: Understanding Machine Learning: From Theory to Algorithms, New York.
- Sharif, Mahmood/Bhagavatula, Sruti/Bauer, Lujo/Reiter, Michael K. 2016: Accessorize to a Crime. Real and Stealthy Attacks on State-of-the-Art Face Recognition, Wien.
- Shaw, Martin 2005: The New Western Way of War. Risk-transfer War and its Crisis in Iraq, Cambridge, UK.
- Stevenson, Beth 2016: Analysis: Taranis Developers Reveal Test Flight Specifics, <https://www.flightglobal.com/news/articles/analysis-taranis-developers-reveal-test-flight-spec-425347/>; 19.07.2019.
- Stocki, Trevor J./Li, Guichong/Japkowicz, Nathalie/Ungar, R. K. 2010: Machine Learning for Radioxenon Event Classification for the Comprehensive Nuclear-Test-Ban Treaty, in: Journal of environmental radioactivity 101: 1, 68–74.
- Sun, Yijun/Li, Jian 2005: Adaptive Learning Approach to Landmine Detection, in: IEEE Transactions on Aerospace and Electronic Systems 41: 3, 1–9.
- Sundaresan, Lalitha/Chandrashekar, S./Jasani, Bhupendra 2017: Discriminating Uranium and Copper Mills Using Satellite Imagery, in: Remote Sensing Applications: Society and Environment: 5, 27–35.
- Szegedy, Christian/Zaremba, Wojciech/Sutskever, Ilya/Bruna, Joan/Erhan, Dumitru/Goodfellow, Ian/Fergus, Rob 2014: Intriguing Properties of Neural Networks, <https://arxiv.org/pdf/1312.6199.pdf>; 19.07.2019.
- Truong, Q. B./Borstad, Gary/Saper, Ron 2005: Integration of Satellite Imagery and Other Tools in Safeguards Information Analysis, https://www.remote-sensing.aslenv.com/documents/ESARDA_paper_2005_TRUONG_etal.pdf; 19.07.2019.
- Tuma, Matthias/Igel, Christian 2009: Kernel-Based Machine Learning Techniques for Hydroacoustic Signal Classification, CTBTO International Scientific Studies Conference, Wien.

- UNIDIR/VERTIC 2003: Coming to Terms with Security. A Handbook on Verification and Compliance. <http://www.unidir.org/files/publications/pdfs/coming-to-terms-with-security-a-handbook-on-verification-and-compliance-en-554.pdf>; 19.07.2019.
- US Navy 2017: MK 15 – Phalanx Close-in Weapons System (CIWS), http://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=487&ct=2; 05.01.2018.
- Yu, Xie/Jing, Meng 2017: China Aims to Outspend the World in Artificial Intelligence, and Xi Jinping Just Green Lit the Plan, <http://www.scmp.com/business/china-business/article/2115935/chinas-xi-jinping-highlights-ai-big-data-and-shared-economy>; 16.11.2017.
- Zeiler, Matthew D./Fergus, Rob 2013: Visualizing and Understanding Convolutional Networks, <https://arxiv.org/pdf/1311.2901.pdf>; 19.07.2019.

PRIF REPORT

PRIF Reports offer background analyses on political events and developments and present research findings.

Fröhlich, Marieke (2019): Masculinities in Peacekeeping. Limits and transformations of UNSCR 1325 in the South African National Defence Force, PRIF Report 7/2019, Frankfurt/M.

Christian, Ben/Coni-Zimmer, Melanie (2019): Deutschland im UN-Sicherheitsrat 2019–2020. Eine Halbzeitbilanz, PRIF Report 6/2019, Frankfurt/M.



www.hsfk.de/PRIF-Reports
www.hsfk.de/HSFK-Reports

PRIF SPOTLIGHT

PRIF Spotlights discuss current political and social issues.

Fehl, Caroline (2020): Syrische Folter vor Gericht. Die partielle Rückkehr des universellen Rechts, PRIF Spotlight 2/2020, Frankfurt/M.

Polianskii, Mikhail/Rogova, Vera (2020): Lost in transition? Putin's strategy for 2024, PRIF Spotlight 1/2020, Frankfurt/M.



www.hsfk.de/PRIF-Spotlights

PRIF BLOG

PRIF Blog presents articles on current political issues and debates that are relevant for peace and conflict research.



<https://blog.prif.org/>

PRIF Reports and PRIF Spotlights are open-access publications and are available for download at www.prif.org. If you wish to receive our publications via email or in print, please contact publikationen@hsfk.de.

www.facebook.com/HSFK.PRIF
www.twitter.com/HSFK_PRIF
<https://blog.prif.org/>

NICO LUECK //

MACHINE LEARNING-POWERED ARTIFICIAL INTELLIGENCE IN ARMS CONTROL

Artificial intelligence (AI), especially AI driven by machine learning, is on everyone's lips. Even in armaments such systems are playing an increasingly important role: Some weapons systems are already able to identify targets independently and engage in combat with them. This poses problems for traditional forms of arms control originally designed to monitor physical objects such as mines and small arms and their internal function. In addition, important additional effects of reliable control such as confidence-building and stabilization of diplomatic relations are not addressed. It is important for arms control to address such risks as well.

At the same time, the deployment of Machine Learning-powered Artificial Intelligence (MLpAI) as a tool offers tremendous potential for improving arms control processes. Here, more precise and comprehensive data processing can engender more trust between states in particular. This tension between the risks and the opportunities connected with the use of MLpAI in arms control is highlighted in this report.