Arms Control for Artificial Intelligence

Thomas Reinhold

Abstract With military weapon systems getting more and more improved by artificial intelligence and states competing about the leading role in this development, the question arises how arms control measures can be applied to decrease this equipment spiral. The ongoing debates on cyber weapons have already highlighted the problems with controlling or limiting digital technologies, not to mention the dual use problems. While still in an early stage, this chapter develops possible approaches for AI arms control by considering the different life cycle steps of a typical AI enabled system, based on lessons learned from other arms control approaches. It will discuss the different starting points, their arms control potential as well as its limitations to provide a holistic perspective for necessary further develops and debates.

1 Introduction: Or why Hard Arms Control for Artificial Intelligence Should Be Considered

In this book, we look at both the possibility of using artificial intelligence (AI) to foster arms control as well as the dark side—the acceleration of warfare or the possible transfer of decision-making from humans to machines. While AI can foster arms control (see the overview by Schörnig in addition to the individual chapters) at the same time it needs to be controlled. In the debates on cyber arms control and autonomous weapons control, confidence-building measures (CBMs) and increased transparency are often seen as the best outcomes. In some circumstances, as in the case of autonomous weapons, only political declarations remain realistic, but their meaning is unclear until they are actually applied. In short: The arguments why hard, verifiable arms control is not possible are varied and compelling, and they dominate the current discourse.

T. Reinhold (✉)
Chair of Science and Technology for Peace and Security (PEASEC), Department of Computer Science, Technical University of Darmstadt, Darmstadt, Germany
e-mail: reinhold@peasec.tu-darmstadt.de

T. Reinhold, N. Schörnig (eds.), *Armament, Arms Control and Artificial Intelligence*, Studies in Peace and Security, https://doi.org/10.1007/978-3-031-11043-6_15

In this text, however, it will be argued that, at least from a theoretical perspective, there are definitely starting points for quantifiable and verifiable arms control measures in the realm of AI and their realization only needs to be consistently tracked and checked. These approaches are very technology-specific, and it is necessary to unpack the concept of AI to start with. Only when AI is broken down into smaller, manageable and technically relevant parts can promising approaches be identified. This makes it necessary to define AI to begin with.

As many chapters in this book have shown, AI is a well-established concept and includes deterministic variants such as expert systems. The current debates among IT specialists, however, focus on so-called *neural networks* and their recent spin-off variant *deep learning* (see the introductory chapter in this book for details; see also Charniak, 2019; Kersting, 2018). Most civilian applications use this latest form of AI, and it is also widely used in the military realm. Consequently, the structural approach adopted here focuses on neural networks but can be applied analogously to many other forms and variants of AI. To provide a broader perspective and avoid restricting the discussion to a specific technological branch, in the following text, the term *artificial intelligence* will be used, including neural networks and earlier, current or even future forms of machine learning (ML).

As will be seen, the development process or *life cycle* of AI based on ML can be broken down into four components with different but always promising approaches being applied to each individual component.

It goes without saying that this text breaks new ground and, also due to its brevity, only formulates initial thoughts. As a result, no silver bullet can be expected, but ideally a crystallization or starting point at which further discussion can begin. This text does not aim at ending the debate but rather at (re)opening it by introducing technical details in order to overcome the common *it won't work because it's complicated* perspective.

To prove that there are more options for applying arms control measures to AI, the text will present arguments as follows: Section 2 will provide a brief overview of current technological trends and the rise of AI and show why it fosters militarization. Section 3 will briefly examine best practices and established arms control instruments in order to describe the variety of options arms controllers can choose from under varying circumstances. After that, Sect. 4 unpacks the development process of AI and identifies the four key components where arms control measures could start. Section 5 delves deeper into this and identifies the best arms control practices for each of the components. Section 6 addresses the problematic field of verification and how the arms control measures suggested in Sect. 5 could be successfully verified. Section 7 debates potential pitfalls and necessary pre-conditions when such ideas and concepts are applied to real-life AI-enabled weapon systems. Section 8 goes back a step and asks whether CBMs could be a viable alternative to the hard and verifiable measures previously suggested. It concludes that some confidence-building could be done, but also argues that confidence-building alone would not be enough given these options. Finally, Sect. 9 summarizes the text and offers a glimpse into the future.

2 The Rise of Artificial Intelligence and Its Militarization

The technology of ML and more generally AI has taken huge steps in recent years, supported by the miniaturization and performance enhancements of IT. Some use cases that traditionally had been a core application area of AI such as visual pattern recognition have been integrated in broadly distributed consumer electronics in the form of facial recognition, image classification or natural speech analysis and its synthesis (see, for example, the text by Schörnig, 2022). Whereas former AI applications usually focused on one specific task and its optimization, the significantly increased amount of processable data has fostered the development of AI systems that become an integral part of complex applications. Such AI systems process, filter and classify huge amounts of data—partly in real time—and are intended to reduce the data overload of many real-world scenarios for human operators—for example high-frequency trading (Briola et al., 2021), social media hate speech detection (Putri et al., 2020) or automated cybersecurity systems (Belani, 2021). And finally, AI is a core element of the ongoing trend toward autonomous systems that are able to navigate under dynamic, partly uncertain or even unknown environmental conditions (see the text by Dahlmann, 2022). These developments highlight a trend in which the role of integrated AI systems is shifting from being one of many subsystems that deliver input or perform dedicated tasks to becoming the core element that integrates all the different subsystems and generates the final output.

These advances are also affecting military trends, applications and strategic decisions as AI seems to provide the core tool for managing the digitalization of military systems and the necessity to process huge amounts of data into machine- or human-usable information (see the texts by Sauer, 2022; Fischer, 2022). Previous chapters analyzed many of these aspects and discussed the problems and challenges that arise from the application of AI for different military technologies such as the automation of cyber defensive and offensive measures (see the text by Reinhold & Reuter, 2022), robotics and autonomous military vehicles (see the text by Dahlmann, 2022), or even the enhancement and automation of nuclear defense systems (see the texts by Heise, 2022; Baldus, 2022). In addition to direct integration into weapons or weapon control systems, ML algorithms are also being inserted into other military applications such as battlefield management, logistics, recruitment and training of personnel, or other aspects of the complex military administration and bureaucracy (Bundeswehr, 2019).

This development is offering new challenges for the regulation, containment, and non-proliferation of AI as a military technology as well as of AI-enabled military (weapon) systems. As the debates over the militarization of cyberspace have already shown, many established measures of arms control and verification are not applicable to digital technologies because of their specific technical features and thus require new methods (Reinhold & Reuter, 2019a). Whereas political measures such as confidence-building, codes of conduct, or norms debates are already taking place (Paoli et al., 2020), technical approaches that would allow verifiable measures

have not yet been studied. Nevertheless, as Lawrence Lessig once stated: "code is law" (1999), pointing out that software and its underlying code directly reflects the rules and values of its creators who set its capabilities and limits. As any AI is based on code, this is certainly true for the military application of ML. As a work of human beings, it can be controlled in principle, shaped and adjusted to serve a good purpose. Used in the right way, AI can supplement potential international norms restricting the military use of AI with actual control, restriction and verification measures.

The following sections offer preliminary thoughts on how this could be done and what obstacles will be faced.

3 Best Practices and Lessons Learned from Other Technologies

As a first step, it is helpful to look at established arms control measures for other technologies in order to understand the lessons learned and the best practices that can potentially be applied to AI. According to Mölling and Neuneck (2001), the different forms of arms control measures that have been developed for chemical, biological, or nuclear (CBN) weapons as well as conventional forces can be roughly broken down into four groups:

- Declarative measures that are based on agreements of Do's and Don'ts
- Usage-related measures and regulation
- Trade and proliferation measures
- Information exchange-based measures

Much like the tools developed for the militarization of cyberspace (Reinhold & Reuter, 2019b), digital technologies lack a direct physical representation apart from the interchangeable storage medium required for usage and proliferation-based measures that try to count or track regulated items. The digital information of AI components can be seamlessly copied, cloned, and distributed, which renders any measure that requires physical objects impossible to apply and complicates verification, but favors declaratory, regulatory and information exchange-based approaches. This does not reduce the value of cooperative measures between states or even possible agreements on trade controls of AI components based on company declarations of the traded goods, but it limits the possibilities for controlling compliance with agreements based on objectifiable information or even monitoring other parties to an agreement without their consent or cooperation.

This aspect highlights the necessity of analyzing the technical foundation and characteristics of AI and its components in order to identify features that can be measured and compared. A similar analysis for cyber tools (Reinhold & Reuter, 2019a) concluded that in addition to the technological challenges that have been mentioned, IT-related products actually do provide quantifiable parameters that could be applied for arms control measures. These include:

- The total power supply as well as the current power consumption of IT infrastructures
- The available supply of cooling systems and their thermal power as well as the current heat production of IT infrastructures
- The available network bandwidth capacities as well as the current flowrate of transmitted data via monitored network connections
- The total extent of connections of monitored networks to other external civil or commercial networks (the so-called *peering*) and their maximum possible transmission performance
- The number of staff required for the maintenance of the IT systems

It is thus possible *in principle* to identify measurable and quantifiable aspects of AI where action to implement arms control measures could be taken. As in the case of cyber tools, the following section will unpack the development process or the *life cycle* of AI to identify where in the development process action could be taken and on which key components.

4 The AI Life Cycle: The Components of Artificial Intelligence Applications

The previous chapters in this book and the many approaches used by the authors alone made it clear that there are and have been many different forms of AI and algorithmic approaches. In the current debates, especially the so-called neural networks and their recent spin-off variant *deep learning* (see the introductory chapter in this book for details) play a major role. These approaches, in combination with the processing power of computers and microchips available nowadays, provide the most powerful results and can mimic human intelligence for the first time, as was envisaged in the early years of this field of research (Charniak, 2019). This dominance has led to the fact that neural-network methods are already used synonymously with the term *artificial intelligence* or *machine learning* in many contexts, even when the technological foundations differ (Kersting, 2018). Consequently, the following structural approach focuses on neural networks, although it can be applied analogously to most of the other forms and variants of AI. To provide a broader perspective and avoid restricting the discussion of arms control to a specific technological field, the following text will use the term AI to include neural networks and previous, current or even future forms of ML. Regardless of the different approaches, all AI applications are marked by a specific *life cycle*: from their development to their deployment. This *life cycle* concept reflects the fact that each AI-enabled application passes through different transformation steps that apply initial algorithm and design decisions in technical software components which are then later combined in the final application.

Facilitating a concept from a report on the security of AI (Stiftung Neue Verantwortung, 2019), the following life cycle illustrates these transformation steps for a military AI application:

1. Definition of the goal and the desired capabilities of the AI
2. Acquisition and preparation of the required training data
3. Choosing the required ML methods
4. Learning the implicit input-to-output rules (the so-called *classifiers*) during training with the selected training data
5. Creating a fully trained AI system (the so-called *model*)
6. Deploying the AI into military systems or effectors
7. Applying the military system or effector, probably with a feedback loop and retraining of the model

In these steps the following four different components are always employed in one way or another as part of any AI application and its development:

A. The training data, that is, the dataset which is given to the algorithm to identify patterns and regularities as well as used for the testing and evaluation of the AI.
B. The classifiers, that is, the representation of the training goal.
C. The model, that is, the final data structure which encompasses the learned interrelations and information.
D. The effector, that is, the actual weapon that achieves the destructive effect under the control of AI.

When debating the chances of implementing arms control measures for AI, it is very important to distinguish systematically between these components, as each of them, together with its associated transformation steps uses different technological approaches and thus provides different technical aspects and characteristics of AI applications that can be used to impose restrictions. This component-centric perspective is useful for maintaining a technical perspective on the possibilities and challenges of AI for arms control. However, while the first three components relate to the development process of the AI algorithm itself, the fourth component is related to its application. As an AI does not directly contain but only controls effectors, it will always be part of a larger military system that provides the actual effectors and must thus be taken into account in arms control measures.

5 The Components of AI Development: Applying Tailored Arms Control Measures

The AI components that have been identified will now be discussed in greater detail, including analysis of the measurable and quantifiable aspects where arms control initiatives can start and where potential technological thresholds between civilian and military AI can be defined. After a review of possible lessons learned from fields

Fig. 1 Data transformation along AI lifecycle development stages. Source: Own illustration

of successful arms control initiatives that might be applied to AI, each subsection that follows will discuss which arms control measures could be applied to each particular component.

5.1 The Training Data

The training data is essential for every AI application that facilitates any kind of learning and adjustment of inner processing capabilities. The data used for training can take many different forms but, in most cases, involves a specific set of information built from streams or batches of raw data and tailored to the specific learning goal as well as the specific variant of learning algorithm. Organization of the data is necessary in order to structure the amount of information presented to an AI algorithm so that it contains enough relevant relationships that can be identified and learned, but does not become too *polluted* with misleading or distracting information. For example, an AI that is required to learn to identify IEDs (improvised explosive devices) in visual information needs to be presented with different images that in the best case contain all different kinds, sizes, shapes, forms of construction, etc., of these devices. A well curated set of training data usually also contains negative data items, in the present example images of devices or objects that are not IEDs. The final curated data set is then usually split into different batches that are used for the training of the AI, for testing the trained model (see Sect. 5.2) with data that has not yet been used for training and which the algorithm has not yet *seen* and a further batch to evaluate the quality of the model. Figure 1 presents an overview of the different stages in the processing pipeline from raw data to applicable training sets as given in an ENISA report (ENISA, 2020).

The different steps in this process can be performed by a single actor or distributed over different institutions or can be provided by commercial vendors or brokers. As the data needs to be collected, processed, and curated in plain text—which means that it cannot be encrypted during this step—it potentially provides options for checking, comparing or verifying against defined principles.[1] This provides the

[1]Experience from civilian applications has shown, however, that datasets struggle with unrecognized biases. If, for example, the dataset scarcely features people of color but focuses on white males, the AI might struggle to recognize black faces (Buolamwini & Gebru, 2018).

following access points for control, regulation, or restriction in possible arms control agreements:

- Restrict the use of specific type of information or limit the scope of raw data collections for specific military training goals, for example, for the identification of human combatants.
- Monitor, regulate or restrict the use of specific raw data aggregation infrastructures (such as dedicated cloud services or sensor systems). In particular, the gathering of training data for possible offensive military applications like autonomous weapon systems (AWS) could be restricted to a certain degree to real-world military scenarios such as actual military operations—which is at least impractical—or to dedicated military testing environments designed for such raw data acquisition. The latter could even be passively monitored.
- As far as dedicated vendors, data brokers or curation services are concerned, their commercial activities could be lawfully regulated, in conjunction with appropriate transparency and compliance control measures. This would provide additional measures for proliferation control.
- In addition to the limitation on use off specific raw data, it is also possible to regulate specific kinds of data curation and preparation that reflect specific, limited or unwanted training goals.

5.2 The Classifiers

The classifiers of an AI algorithm represent the training goal and in the case of successful training the application quality of the AI system. The exact shape of the classifiers depends strongly on the AI algorithm that is used, but they always reveal the intent of the trainer. Although regulation of this kind of thing is always a challenge for arms control, the following approaches are possible.

- Limit or prohibit the usage of specific types, ranges or characteristics of classifications in order to limit the application scope of AI systems.
- Intentionally limit classifier quality in order to reduce the applicability and degree of autonomy of AI systems and thus enforce closer human interaction and a wider decision range, for example by allowing the classifier to identify humans but not to provide an assessment of their combatant status, leaving this to human judgment.

However, it is not the aim of arms control to check used datasets for biases but to prevent the use of certain datasets which could be used for undesired weapon systems.

5.3 The Model

The model of an AI system, as the trained state of an AI that is ready for application, is the embodiment of the intended goals implicit to the training data and the selected classifiers. With regard to the AI system itself, the model is the final product that could be built into the designated external system which it supports or controls. Whereas some models could be highly specific for a designated use case and external system, others could be more *off the shelf*, generalized and applicable to a huge variety of external applications. Thus, the following possible arms control measures exist:

- Regulation or restriction of the proliferation of models trained for specific military purposes such as distinguishing between civilians and human combatants.
- Control or prohibition of trade in models in conjunction with Wassenaar-like information and transparency measures.
- Restriction of the use of specifically trained models, either for direct application in an external system or for use as the basis of further AI training scenarios.

The ongoing trend to miniaturization and specialization of microchips that provide among other things AI-optimized hardware also requires regulation. But since such hardware parts are not designed for a specific use case but rather to be equipped with a dedicated AI model, their regulation raises strong dual-use concerns, as will be discussed below.

5.4 The Effectors

An AI-enabled application has—in contrast with most other militarily weaponized technology—no direct effect on its environment. Instead, the AI will always be part of larger weapon systems in which it controls specific aspects of the system or controls it completely up to the release of the actual effector—tasks that mostly had been or are still assigned to human operators. In many cases the weapon system itself is not a new development and the AI is simply an extension or upgrade, enhancing systems like air defense, uncrewed vehicles, battlefield command and control, or cyber defense measures. This means that in the best-case scenario, these weapon systems are already part of arms control agreements that can be adapted to include AI-specific regulations. A second aspect of this relationship is that it directly relates to the question of the limitations and boundaries of the autonomy of weapon systems or trigger decisions and the debates about control of the acceptable extent of such capabilities. In conclusion, the following arms control measures are applicable:

- Extend existing arms control treaties to include the enhancement or replacement of components of the regulated items and technology with AI applications or include these aspects in negotiations on the renewal of terminated treaties.

- Include AI applications or systems that are intended to be integrated into weapon systems in existing arms trade and non-proliferation agreements such as the Wassenaar Arrangement (WA).
- Expand discussions, international security debates, and treaty negotiations on the regulation of AWS to encompass various aspects of and potential integration of AI and its consequences—including its regulation according to the International Humanitarian Law (IHL).

6 Verification

The previous section has shown that there are indeed starting points for applying hard arms control measures to AI if the dazzling term *AI* is broken down into technically elementary components. However, probably the hardest part of applying arms control measures for digital goods is the challenge of how compliance with agreements can be verified (see the text by Schörnig, 2022). This also applies to AI algorithms and the situation is additionally complicated by the black box character of an AI (see the text by Verbruggen, 2022). This technical aspect arises from the fact that the model of an AI does not provide a human-readable or comprehensible representation of the learned states and the algorithmic micro-decisions it makes. For current AI algorithms, all that can be seen is the output arising from a given input, not the path that led to this conclusion. This raises the question of which parts of an AI application could be controlled in terms of defined thresholds or prohibitions. The following list presents initial ideas for dealing with this challenge. It is not meant to be complete and is highly dynamic in view of emerging technical developments in the field of AI.

- Training a clean model with the data that was allegedly used for the original AI must create a model that works identically to and generates the same results as the defined set of testing input. This makes it possible to verify whether an AI has been trained with a set of training data that complies with agree-upon rules. This method is limited to static AI applications that are not re-trained or otherwise adapted during their real-life application, as adaptation changes their internal state and thus undermines comparability.
- To verify that decisions made by an AI comply with certain rules, it is possible to use a set of test data specifically constructed to contain triggering input which will lead to a specific output. As a trivial example, an AI could be trained to identify tanks in images and tag them as military targets under the restriction that it will not tag other objects or even humans as targets and will untag tanks that are relatively close to humans. A test set of images would include images of tanks as well as humans in different surroundings and combinations. Tested against these images, the AI must only tag the tanks that are not surrounded by humans.
- Newer technological developments of specific AI algorithms may provide the technical means for the verification of decisions. A research trend involving

so-called *explainable AI* (Vilone & Longo, 2020) provides a retraceable input-to-output path which at least makes it possible to understand the technical process that resulted in a particular decision and permits conclusions on the influence of the training on the real-world performance that followed it. Even if this procedure is not capable of identifying the effects of specific training input, it can provide understanding of how specific clusters of training data modified the final model and its data processing. Since such algorithms require the storage of additional information as well as the necessary data processing, such features usually reduce the overall performance of the AI algorithm. As it would unravel the black box character as an important precondition for arms control, explainable AI can provide an important tool, but will have to be made mandatory in agreements in order to be implemented.

- A final challenge lies in the task of verifying whether an AI has been used as part of an existing system without deep analysis of the operation system. In most cases this is not accepted by treaty members. This resembles the often-cited and still valid idea of the Turing test in connection with the choice between an AI and some other form of deterministic algorithm with hard-wired instructions. Under optimal conditions the latter will always provide an output that is predictable, as it can be calculated externally as long as the hard-wired instructions are known. An AI on the other hand is designed to provide the best possible approximations to the exact result for an input that has not been used during the training phase. These differences between the actual and the exact result might be used to identify the application of an AI.

In the military there is a saying i*t takes one to get one* meaning that in certain situations symmetry is the only possible response or is needed to counter a specific capability. This poses the interesting challenge of using an AI to verify other AI. As the algorithms involved in ML are—in addition to other uses—perfectly applicable to detecting patterns within unknown data or separating and classifying complex information, it is at least theoretically possible to train a *verification AI* with the output from another AI that needs to be monitored. The results yielded by the *verification AI* could then make it possible to draw conclusions concerning the learned processing rules of the AI being checked or the training input it is assumed to have received. Even if such thinking is futuristic at present and applicable measures have yet to be developed, it could serve as the basis for establishing measures for controlling compliance with agreed rules.

7 Pre-conditions and Pitfalls for Arms Control

Many of the ideas discussed and considered above are currently still no more than theory and appropriate technical approaches need to be developed, tested, and—conceivably if ever—installed as measures for arms control. This is, on the one hand, a direct result of the fact that AI is a relatively new topic in military technology,

whose capabilities have been boosted by the development of more efficient algorithms alongside dedicated hardware. On the other hand, its implications and limitations are not as yet fully understood and arms control measures for AI technology must consider the specific conditions discussed above.

The first of these is the obvious and presumably most influential issue of the highly dual-use character of AI and ML. As AI and its components are inherently only parts that are included in more comprehensive systems for specific tasks, the regulation of explicitly military AI will turn out to be inefficient. Although AI applications that are specifically trained with military-grade information and intended to cover specifically military use cases presumably either already exist or will eventually do so, in most cases more generic AI components will be produced and acquire capacities (such as image recognition, information clustering, etc.) that will later only need to be adapted to specific tasks. This aspect also relates to AI-specific hardware that is experiencing strong demand and a corresponding driving force in civil commercial products such as consumer electronics. The further miniaturization of such generically applicable technology will probably further strengthen a trend toward cheap off-the-shelf hardware that is ready to be deployed.

A further aspect relates to the current technological imbalance of AI technology. Although a great deal of groundwork has been carried out in recent decades and published in scientific journals and conference proceedings, the current trend in the implementation of AI in real products is being driven by a small number of technological global players that hold the intellectual property rights. It is foreseeable that these companies, and with them the states where they operate, will try to defend this head start in order to preserve the advantages they have gained from this technology in both commercial and also military domains. This imbalance between the *haves* and the *have-nots* will probably complicate the establishment of arms control measures as it has to deal with inherently opposing interests. In addition, AI research and its development have a strong dual-use character. As the actual use of an AI is primarily determined by its training, the underlying algorithms involved in how exactly the model is developed on the basis of input information or how classifiers are created and applied is the same for military as well as civil uses and application. This aspect also includes dedicated AI hardware such as specific microchips that are optimized to perform the required AI calculations or feature a specific technical design that is adjusted to AI models such as neural networks. This complicates the regulation of AI algorithms and their implementation in specific hardware.

Another issue relates to the problems that have already been discussed regarding the technical challenges involved in verifying AI arms control measures. The characteristics of digital goods provide many chances and opportunities for hiding non-compliant behavior while simultaneously hindering effective control mechanisms. In addition, the availability of related commercial products makes it easier to establish a dedicated domestic industry for military-grade AI. This might either prevent states from joining such *toothless* agreements or—on the contrary—might even offer states an incentive to dishonestly sign treaties safe in the knowledge that

non-compliance is not trackable. This challenge may be eased with further technological developments but so far is a game stopper.

The final aspect that will probably hinder the establishment of arms control measures for AI concerns is the perception of this technology as mostly unproblematic and not dangerous enough. In most proposals, research projects or statements from military decision-makers, AI is seen as an enabler for military systems or as an enhancement for human tasks. Although debates in other areas such as lethal autonomous weapon systems (LAWS) discuss the threats and problems that arise from decisions made autonomously by machines, these concerns have so far not been included in AI debates to a sufficient degree. As long as AI is not perceived as another aspect of the same problem, there will not be sufficient incentives for states to debate its regulation and the limitation of its military application.

8 Confidence-Building Measures for Military AI Applications: An Alternative?

The preceding sections have shown that the application of *hard* and verifiable arms control measures is not impossible. But just starting to think about the possibilities requires extensive technical knowledge—knowledge that arms control experts often do not possess. Consequently, the first step toward actual arms control agreements has often been the establishment of confidence via CBMs. In most cases this step has involved, among other things, the exchange of information about national security interests and concerns about shifting military capabilities resulting from technological developments as well as technical details of new developments. These measures for achieving transparency are intended to allow potential adversaries to gain an understanding of the military impact of the adoption of new technological developments as well as of their limitations. With regard to the influence of AI on military developments, the following details of the different components of an AI could be made available as part of CBMs in order to understand its impact:

- Samples of the training data related to the intended capability of the AI
- Training environments or data aggregations sources
- The classifiers and the features that are intended to be detected and processed for the output of the AI
- Details of the application of the AI and the facilitation of its output with regard to the complexity and the degree of freedom that the AI's decisions are used for
- Details of the system that the AI is part of (e.g., effectors, military relevance, and facilitation)
- Information on the structural changes in tactics or on organizational changes where AIs are used to enhance human decisions or replace them

Regarding the similarities that AI shares with other digital technologies, it is important to highlight the contrasting conceptualizations of AI in existing debates on CBMs for cybersecurity and cyberspace in international forums like the Organization for Economic Co-operation and Development (OECD) or the United Nations (UN). As military capabilities are mostly shaped by human skills in cooperation with intelligence-gathering operations, the debates on cyber CBMs mostly cover its impact on military defensive and offensive strategies, but seldom involve technical details or technological knowledge. On the other hand, the military development and application of AI is driven far more strongly by active scientific research on AI algorithms and dedicated hardware and is thus influenced by issues of intellectual property and maintaining a technological edge in knowledge. Thus, although it might be meaningful to promote existing cyber forums on CBMs, these debates will probably face greater reluctance by participating parties to share the technical details mentioned above and may have to focus more on strategic goals.

9 Conclusion: Or How AI May Develop and What Arms Control Can Do About It

When looking at current trends in AI, it is safe to conclude that one way or another AI will find its way into military applications. Even if the current level of attention is reduced or has to face the inherent limitations of this technology, the normative power of the factual as well as the money currently being spent on Research and Development (R&D) will bring the world AI-enabled military systems. This will probably happen regardless of whether they actually perform better, as long as they promise to shorten the sensor-to-trigger loop or otherwise seem to supersede human cognition and reaction limitations. On the other hand, it is doubtful that we will see any kind of an envisage complex AI systems that integrates and controls complex battlefield activities in the near future because the complexity of such activities conflicts with the single-purpose performance of AI algorithms. At best, there will be an integration of multiple specific AI applications, each optimized and facilitated for a dedicated task that will be integrated into such systems, much as is already the case for self-driven cars that consist of multiple interoperating AI applications. Another issue is the currently strongly divided technology ownership. It is quite possible that, regardless of its actual usage, the most advanced AI countries will continue to perfect AI capabilities or even further extend them in order to maintain their current advantage. This could result in strategic benefit or be at least a bargaining chip in international power struggles. In addition, as AI is—in contrast, to for example the cyberspace area—strongly connected with intellectual property rights and technological research and knowledge, this will probably be closely accompanied by economic and trade restrictions. As AI hardware becomes more and more important and a question of performance, such issues could even spill over to the current international disputes and struggles to create national sovereignty over microchip

design and production (Kleinhans & Baisakova, 2020). From the standpoint of military technology, the ongoing trend toward miniaturization of computation devices that also includes AI hardware may foster and accelerate a shift from current military R&D projects involving large monolithic AI systems for complex tasks to the integration of dedicated AI capabilities into small military systems and consumables such as small arms, land mines and ammunition. As small arms are still the real weapon of mass destruction (WMD), AI-enabled small arms with self-guiding ammunition might be even more terrifying and deadly.

This leaves a great deal of work for further arms control approaches and requires substantial convincing of national and international actors. It probably also means that in the near future AI will become one of the many factors that need to be discussed and considered in connection with many existing weapon systems and military capabilities. This could also increase the necessity of including AI in existing arms control treaties. Measures for AI face issues similar to those involved in the militarization of cyberspace, where many established arms control approaches have not worked and have thus led to a need for new technical solutions and tools for verification. As AI and cyberspace share a great deal of underlying technology it probably makes sense to combine discussions and the development of arms control tools based on these technologies. On the other hand, AI-enabled applications or military systems will still rely on small-scale single-problem AI solutions so that there will still be opportunity for approaches to its regulation that focus on specific details, technical features, or capabilities, without the necessity of tackling the sci-fi vision of a *super-AI*. This also means that verification measures—despite the problems mentioned—could be built upon very detailed features, which, from a technical perspective, leaves room for optimism. And that is something that arms control has always needed.

References

Baldus, J. (2022). Doomsday machines? Nukes, nuclear verification and artificial intelligence. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.

Belani, G. (2021). The use of artificial intelligence in cybersecurity: A review. *IEEE Computer Society*. https://www.computer.org/publications/tech-news/trends/the-use-of-artificial-intelligence-in-cybersecurity

Briola, A., Turiel, J., Marcaccioli, R., & Aste, T. (2021). Deep reinforcement learning for active high frequency trading. *arXiv*. https://doi.org/10.48550/arXiv.2101.07107

Bundeswehr. (2019). *Künstliche Intelligenz in den Landstreitkräften.* https://www.bundeswehr.de/de/organisation/heer/aktuelles/kuenstliche-intelligenz-in-den-landstreitkraeften-156226

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research, 81 (pp. 1–15).* 2018 Conference on fairness, accountability, and transparency.

Charniak, E. (2019). *Introduction to deep learning*. MIT. https://doi.org/10.5555/3351847

Dahlmann, A. (2022). Armament, arms control and artificial intelligence: The impact of software, machine learning and artificial intelligence on armament and arms control. In T. Reinhold &

N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.

ENISA. (2020). *Artificial Intelligence Cybersecurity Challenges - Threat Landscape for Artificial Intelligence*. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

Fischer, S.-C. (2022). Military AI applications: A cross-country comparison of emerging capabilities. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.

Heise, A. (2022). AI, WMD and arms control III: The case of nuclear testing. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.

Kersting, K. (2018). Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines *Frontiers in Big Data, 1*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931929/

Kleinhans, J.-P., & Baisakova, N. (2020). *The global semiconductor value chain: A technology primer for policy makers.* Stiftung Neue Verantwortung. https://www.stiftung-nv.de/sites/default/files/the_global_semiconductor_value_chain.pdf

Lessig, L. (1999). *Code and other laws of cyberspace*. Basic Books, Inc.

Mölling, C., & Neuneck, G. (2001). Präventive Rüstungskontrolle und Information Warfare. In *Rüstungskontrolle im Cyberspace. Perspekt. der Friedenspolitik im Zeitalter von Comput. Dokumentation einer Int.* Konf. *der Heinrich-Böll-Stiftung am 29./30. Juni 2001 Berlin* (pp. 47–53).

Persi Paoli, G., Vignard, K., Danks, D., & Meyer, P. (2020). *Modernizing arms control: Exploring responses to the use of AI in military decision-making*. UNIDIR. https://unidir.org/publication/modernizing-arms-control

Putri, T. T. A., Sriadhi, S., Sari, R. D., Rahmadani, R., & Hutahaean, H. D. (2020). A comparison of classification algorithms for hate speech detection. *IOP Conference Series: Materials Science and Engineering, 830*(3). https://iopscience.iop.org/volume/1757-899X/830

Reinhold, T., & Reuter, C. (2019a). Arms control and its applicability to cyberspace. In C. Reuter (Ed.), *Information Technology for Peace and Security - IT-applications and infrastructures in conflicts, crises, war, and peace* (pp. 207–231). Springer Fachmedien Wiesbaden.

Reinhold, T., & Reuter, C. (2019b). Verification in cyberspace. In C. Reuter (Ed.), *Information Technology for Peace and Security - IT-applications and infrastructures in conflicts, crises, war, and peace* (pp. 257–275). Springer Fachmedien Wiesbaden.

Reinhold, T., & Reuter, C. (2022). Cyber weapons and Artificial Intelligence – Impact, influence and the challenges for arms control. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.

Sauer, F. (2022). The military rationale for AI. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.

Schörnig, N. (2022). Introduction. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.

Stiftung Neue Verantwortung. (2019). *Securing artificial intelligence*. https://www.stiftung-nv.de/sites/default/files/securing_artificial_intelligence.pdf

Verbruggen, M. (2022). No, not that verification: Challenges posed by testing, evaluation, validation and verification of artificial intelligence in weapon systems. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence: The Janus-faced nature of machine learning in the military realm*. Springer.

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: A systematic review. *arXiv*. https://doi.org/10.48550/arXiv.2006.00093