Fraunhofer

VVS

FRAUNHOFER GROUP FOR DEFENSE AND SECURITY

POSITION PAPER

RISE OF ARTIFICIAL INTELLIGENCE IN MILITARY WEAPONS SYSTEMS

RISE OF ARTIFICIAL INTELLIGENCE IN MILITARY WEAPONS SYSTEMS

THE NEED FOR CONCEPTS AND REGULATIONS

Abstract

The last few years have seen a dramatic increase in the capabilities of artificial intelligence (AI) systems, introducing new risks and potential benefits at the same time. In the military context, these are discussed as enablers of a new generation of »autonomous« weapons systems and the related concept of a future »hyper-war«. Especially in Germany, these ideas are facing a controversial discussion within society and politics. Due to the worldwide increasing use of AI in some sensitive areas such as in defence an international prohibition or a legally binding instrument on the issue is not realistic.

Before deciding on specific policies, a shared understanding of the risks and benefits of this technology has to be gained, including the reaffirmation of fundamental ethics and principles. The application of lethal forces must be directed and controlled by a human, for only humans can be held accountable. The Bundeswehr is aware of the need to deal with these developments in order to be able to fulfil its constitutional mission of defending the country in all future scenarios and against adversaries employing such systems to act in accordance with their development plans. Therefore, the need for concepts and legally binding regulations aimed at **controlling the risks while accessing the benefits** is urgent.

This position paper explains the view of Fraunhofer VVS regarding the current state of the art, explores benefits and risks, and presents a **framing concept for explainable and controllable AI**. Selected research topics necessary to implement the presented concepts are identified and discussed, outlining a path to **trustworthy AI** and the **responsible usage of these systems in the future**. The implementation of the concepts and **regulations** following a **reference architecture** is the **key enabler for acceptability of AI-based weapons systems,** a prerequisite for acceptance.

1. INTRODUCTION – CURRENT SITUATION AND SCOPE OF THIS POSITION PAPER

The importance of artificial intelligence (AI) has dramatically increased over the course of the last years. While the **ethical dimension of AI** in the **civilian domain** has been covered extensively, there is a lack of sober and in-depth analysis of AI in the military domain [1]. Nevertheless, it is one of the most controversially discussed aspects in the usage of AI methods, in particular as a component in weapons systems. A global ban on AI-based weapons systems is unlikely. Therefore, the Bundeswehr is forced to deal with such systems according to their constitutional task. AI is a complex field of research and concerns are common that it will change the nature of armed conflicts for the worse by introducing unpredictable risks. However, it also appears to be an opportunity to increase the precision and scalability of weapons effects – leading to a potential minimisation of unnecessary damage or casualties, saving human lives and resources. See [2] for a more general discussion.

This position paper aims at two points. First, to set the **scope of Fraunhofer VVS**, and second, to define the most **pressing aspects for research and development necessary** in this scope in order to guarantee a **responsible, safe and controlled usage** of this technology as well as the **transparent explainability of the behaviour** of these systems. The implementation of these aspects is fundamental for the acceptability of AI-based weapons systems.

The challenge to scientists, the defence industry, politicians and society in Germany is to find and agree on an approach that allows the fielding of systems containing AI components in order to match the response time and effectivity of weapons systems currently developed in many countries, while still having **effective and reliable ethical and legal control mechanisms**. More precisely, it must be guaranteed that the actions of the system are **compliant** with the **UN Charter** and the respective rules of engagement. This raises the question of **trustworthiness** and how the **compliance and predictability** of AI systems can be assured under all circumstances.

Public opinion in the western world and especially Europe is still divided on the question if an **artificial system should ever be in the position to autonomously decide if a weapons system should be used against a – possibly human – target**. This leads to one of the most important distinctions of AI methods: between »automatic« and »autonomous« systems.

The position of Fraunhofer VVS is to discourage autonomous systems that target humans directly.

In this position paper, research topics and necessary steps **will be focused on automatic systems for physical weapons**. However, this constraint does not limit the scope of the application. AI systems of all degrees of complexity can be a vital part of a complex system of systems, encompassing a variety of different sensor and active technical systems as well as humans. Also, the application range stretches from the tactical level of decisions and proposed actions in the scope of a constrained engagement up to a strategical level, aiming at situational awareness and complex decision support.

This leads to the central question of responsibility. One of the most controversially debated points in current discussions is the **level of independent action** an AI system may put into effect. The more **static rules and limited freedom an AI system has, the more predictability and consistency** the results contain. On the other hand, more freedom for the AI system can increase its versatility and usefulness, but also **increases the amount of responsibility the human-in-the-loop** must take in the system and its decisions. Not being a person per se, AI systems cannot have responsibility or held accountable after an incident in the sense these concepts are usually understood by the public. Since AI systems are programmed and employed by humans, the consequences and legal responsibilities must be attributed to humans. Here, neither the optimal trade-off nor the best practice is clear and needs further research.

A special question is the aspect of **qualification and testing** of given AI systems. The current state of the art does not provide a definition of AI-specific test procedures to **build trustworthiness and confirm predictability**.

This leads to the **conflict between the developers and manufacturers** of weaponised AI systems on the one side, the operators (the military) on the other side, and the judiciary supported by the research community in the middle.

The Fraunhofer VVS aims to outline the concept of trustworthy AI and provide impetus on the necessary steps in research and development to ensure responsible, safe and controlled usage of AI in accordance with fundamental ethical and legal rules in military weapons systems.

2. INTELLIGENT SYSTEMS

»Intelligence« and »autonomy« are omnipresent phenomena in all aspects of life. Before any scientific reflection or technical realisation, living beings fuse sensory impressions with information they have learnt themselves and with communications by other living beings. In this way, they perceive relevant aspects of their respective environment in order to act in accordance with the objective and context on this basis.

AI or digitalisation in a broader sense are the transfer of this concept to the assisting infrastructure and IT structure designed to help humans in their behaviour to understand their surroundings and to react accordingly. As such, digitalisation in the field of defence technology assists the military user on a high level in the areas of command, reconnaissance, impact and support. These abilities will expand the Bundeswehr's capability profile in a significant way [3]. It has to be considered that the estimation of the capabilities of both its own and its adversaries' AI-based weapons systems will become more difficult.

A unique definition of an intelligent system in computer science is just as difficult as the definition of human intelligence in psychology. Hence, in this article, we shall use the term »intelligent system« in a very broad sense, thus encompassing a wide domain of AI-based software systems, which exhibit goal-oriented, highly automated behaviour in complex dynamic environments. One technical solution are systems following an agent-based paradigm, being situated in the environment they sense, fuse, reason and act upon in a proactive fashion [4]. The class of cognitive assistant systems improves the situational picture and awareness of a human operator, using machine learning, data mining or knowledge-based expert systems. Intelligent systems already compete with and are beginning to outmatch humans in some higher cognitive capabilities, such as semantic understanding, reasoning and decision-making. Two automation paradigms are paramount in human–machine interaction, as shown in figure 1 [5]. In supervisory control mode, the intelligent system is tasked and monitored by a human operator and directly interacts with the environment. An assistant system on the other hand receives sensory data and creates a situational picture that is used by the operator for decision-making.

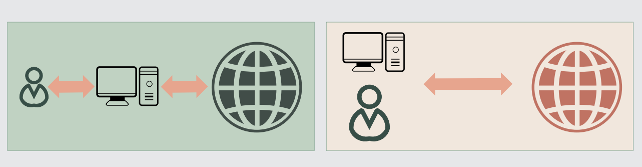

Figure 1: Supervisory control and assistant system model of human interaction with the intelligent system.

Therefore, artificially intelligent support systems that are automated at various levels are relevant for all dimensions of the operational areas (land, air, sea and space, as well as the cyber and information space). One could speak of assisted perception and action in the increasingly complex techno sphere in which military operations are to be carried out. For this reason, the digital transformation is becoming a key to »information, command and impact superiority« as well as to »improving their ability to act and respond« [6] [7]. Since potential adversaries also use or will use innovative digital technologies, spatial delimitation and temporal acceleration will characterise future military conflicts in which cause-and-effect chains will become increasingly automated. Even in the 1950s when the term AI was first coined, the Bundeswehr was aware of this problem [8] [9] [10]. Therefore, everything must be done to »make people in situations that challenge their responsibility experience the consequences of their actions and omissions« [9].

If it is accepted that the Bundeswehr must be able to act on an equal footing with adversaries **under any condition**, defence digitalisation must not be limited to the areas of reconnaissance and enhancement of effectiveness but must also guarantee its technical and ethical controllability and support responsible use [11] [12]. This principle has been part of the Bundeswehr's identity from the very beginning:

»The more lethal and far-reaching the effect of weapons becomes, the more necessary it becomes that people behind the weapons know what they are doing. Without the commitment to the moral realms, the soldier threatens to become a mere functionary of violence and a manager« [9].

In line with its identity, the Bundeswehr's digital transformation is aimed at the technical support of the perceptive reason and the active will of those persons who are responsible for their perception and action. The concepts reason, will and responsibility bring fundamental ideas about the human being as a person into view that imply ethical dimensions. In order to avoid tying our considerations to certain schools of thought in the interest of the broadest possible consensus, these concepts, despite their diverse philosophical connotations, are not outlined more precisely, but are linked to the areas of leadership, reconnaissance, impact and support.

Examples of the application of intelligent systems in security and defence settings include robotic planetary rovers or autonomous underwater vehicles that sense, decide and act in a self-sufficient remote environment. Automatic face recognition systems have been used and are controversially discussed for civil security applications. Automated target recognition algorithms are already in use in current military weapons systems and rely upon similar pattern recognition methods. The upcoming next generation of networked weapons systems, such as FCAS (future combat air system) and MGCS (main ground combat system), will heavily rely on AI methods to process the vast amount of sensory data, reduce operator workload, and improve decision-making in complex dynamic scenarios.

The discussion in the roll-out of many of these examples shows that since digital technologies as a whole, but above all their military use, are accompanied by diffuse scepticism, false expectations and fears that are not always well founded, the scientists commissioned to research them have a particular duty to **clarify the ethical problems arising from their military use**. For this reason, selected aspects are discussed from an information science and engineering science perspective. The administrative action supporting digitisation in the portfolio of the Federal Ministry of Defence must be taken into account. Especially in military use the **basic problems of digital technologies become clear**. As mentioned above, regulations for AI´s trustworthiness are considered as a safeguard and an enabler. **If this problems are solved here, new paths will open up for civil use as well.**

3. BENEFITS AND RISKS

The ever-increasing digitalisation of the battlefield is creating highly dynamic scenarios, which are currently being discussed by NATO under the term »hyper-war«. This involves the combination of classic battlefield elements with attacks in the cyber and information domain and the deployment of large amounts of automatically and autonomously controlled, even unmanned, systems. Due to the significantly increased dynamics of this battle, the term »fight at machine speed« is also often used. It has always been the goal of the military to gain information superiority, to derive command superiority from this and ultimately to achieve combat superiority. It is all about being militarily superior and able to plan and execute operations more precisely and faster than the opponent. The basic principle behind this also translates to civil fields of application like the finance or the security sector.

At a certain point in time, however, humans are no longer able to independently lead this highly dynamic battle in all its details. Technical support in monitoring and evaluating the situation, planning the action, and finally executing the operation will become a necessity. This is where artificial intelligence methods come into play.

The use of artificial intelligence offers the **potential to secure military superiority** in many areas. It begins with situation monitoring and assessment. Due to the digitalisation of the battlefield, more and more information will be available for situation assessment in the future. Pattern recognition methods can lead to these processes being carried out much faster and more precisely. In combination with methods for targeting, target acquisition and fire control, the sensor-to-shooter loop can be accelerated significantly. An improvement in the precision of the effect can lead to a reduction in collateral damage and thus protect the civilian population. By using artificial intelligence, future battles can be conducted much faster, more precisely and more cost-effectively. In addition, technologies such as swarms of drones offer a military capability that does not exist in this form today. A large swarm of drones is a great example which shows that at one point a threat cannot be fought by having a human-in-the-loop who selects targets individually to combat them. In addition, the emergent behaviour of self-organised AI systems must be carefully examined, analogous to natural swarm behaviour.

However, the use of technology from the field of artificial intelligence also brings risks and challenges. Firstly, there are risks that also exist in the civilian use of AI, such as questions of **fairness and impartiality, lack of explainability, or vulnerability against manipulation and misuse**. In addition, a special challenge arises in the military environment when AI technology is used in weapons systems. This **results primarily from the legal classification of the use of weapons systems**. For some years now, there has been an international discussion about the prohibition of lethal autonomous weapons systems (LAWS), especially in the context of the UN convention on certain conventional weapons (CCW). There is a constant debate about what makes a weapons system lethal and what makes it an autonomous weapons system and where exactly the boundaries are. Especially the lethality of a weapons system has an influence on how its use is to be legally assessed. Currently, **a distinction is being made here as to whether the system is designed to be used specifically against humans**. How the discussion will develop cannot be foreseen at the present time.

International humanitarian law (IHL) defines three important principles which have to be considered when **using weapons** in conflicts, besides imputability and precaution aspects. These are the **»distinction between the civilian population and war party«**, the weighing of the **»proportionality of means«** and the assessment of the **»military necessity«** of the use of weapons. This leads to the notion that humans must be able to execute »effective control« when using a weapon. Effective control here means that **the human must be able to understand and judge the entire situation**. Therefore, a simple yes/no decision based on a proposal of an AI system does not correspond to the general understanding of »effective control«.

In addition to the need for technical research to develop suitable procedures for the use of AI technology in the context of the digitalisation of the battlefield, there are challenges that have not yet been solved, particularly with regard to the **aspect of effective control by humans** in the use of weapons systems. The question arises of how to ensure that effective control can actually be executed. What information must be available and how must it be presented? In which situations can **humans be in-the-loop** when using weapons? Which considerations with regard to IHL can be made before a concrete battle start and which are strongly dependent on the course of the battle and the current situation? It has to be examined to what extent **regulations such as international humanitarian law, international law of war or specific regulations can be technically represented and considered by AI technology**. Overlying all of this is a basic need for trust and therefore a concept of how to establish trustworthiness on all levels of this highly complex field of application.

4. BASIC CONCEPT FOR TRUSTWORTHY AI IN FUTURE MILITARY WEAPONS SYSTEMS

A concept for trustworthy AI in future military weapons systems has to consider different phases along the life cycle of an AI weapons system. All phases need to follow a **regulation flow to ensure trustworthiness and the responsible use of AI**. The reference architecture (explained in detail in this chapter) describes affected components and the role of different stakeholders focusing on the AI capabilities. Key enablers for utilising AI in weapons systems are common language, embedded goal, target and effect analysis and assurance of trustworthiness.

Trustworthiness is an important property for any system, and for those associated with significant risks and tightly interacting with human operators this is particularly true. Due to the inherent characteristics of AI components, especially when machine learning is concerned, it can, however, be **technically difficult to guarantee important properties of trustworthiness, like full comprehensibility**. Moreover, trustworthiness is not only about the technical properties of individual AI components but about the properties of the complete system in its overall context. As the ethics guidelines for trustworthy AI put it: »Striving towards trustworthy AI hence concerns not only the trustworthiness of the AI-system itself, but requires a holistic and systemic approach, encompassing the trustworthiness of all actors and processes that are part of the system's socio-technical context throughout its entire life cycle« [13].

The guidelines further provide a good general definition of trustworthy AI based on three components that should be met throughout the system's life cycle.

Trustworthy AI
- should be lawful, complying with all applicable laws and regulations;
- should be **ethical, ensuring adherence to ethical principles** and values; and
- should be robust, both from a **technical and social perspective**, since, even with good intentions, AI systems can cause unintentional harm.

Each of these components implies certain challenges which need to be tackled:

- **Compliance with laws and regulations** is difficult due to the international differences and the lack of strong international guidance, although the law of nations provides a star ting point.
- General **ethical principles** exist, but not so much in the form of internationally recognised guidelines for systems (with or without AI).
- **Robustness**, here certainly also comprising aspects such as **safety and security**, is a more technical dimension which presently also comes with many challenges. Here, future research is required, as outlined later in this document.

Reference architecture

To force the development of trustworthy and reliable AI components, a reference architecture is presented below. It comprises four phases: development, governance, mission preparation and deployment. These four phases are intended to cover the entire process from the development of military AI systems to their use in order to ensure traceability. Regulation, adaptation and feedback steps are inherent in the process. Each phase requires the cooperation of different institutions and actors.

The first phase is the **development phase**. Here the cooperation of all relevant stakeholders is required; in addition to the developers themselves, politicians, military and judicial authorities are needed to establish a legal framework. Ethicists and laywers play an important role in this phase and must define **the ethical and legal responsibility** of an AI system.

To **ensure trustworthiness**, adequate processes, methods and techniques have to be used and a corresponding culture (similar to safety culture) must be established in the companies. In general, we deem it advisable to use approaches similar to those known from safety engineering to ensure the key properties of military weapons systems with AI components.

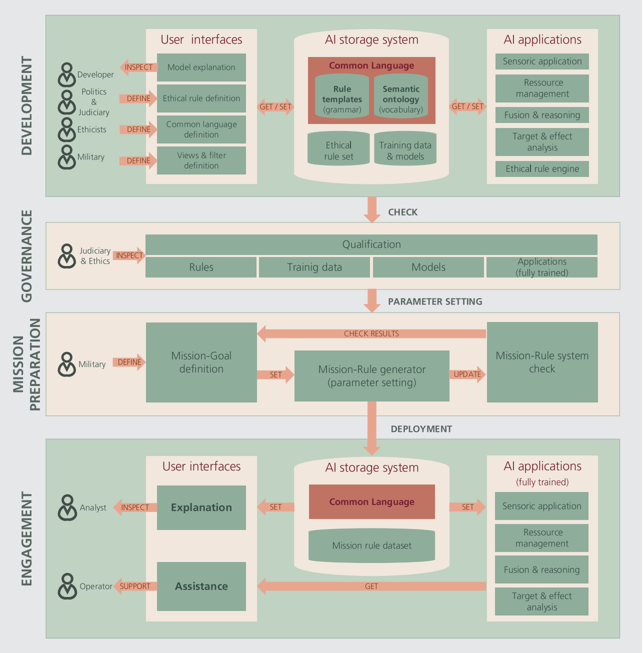

Figure 2: Phase model of an AI-based weapons system.

The key requirements (particularly those related to trustworthiness) of the system under development are analysed, thoroughly specified and broken down onto the system architecture. Further analyses are conducted, and means are identified and integrated into the system to ensure the fulfilment of the requirements. A comprehensive argumentation is built on how the final system ensures the requirements, and evidence is generated (e.g. by testing) to bolster this argumentation.
The results of these activities can be organised as a trustworthiness assurance case, which can then be the basis for a thorough qualification involving the assurance case itself as well as any other relevant development item as linked by the

assurance case. This qualification is to be conducted – like every qualification in the military context – by a sovereign institution with personnel trained in law and ethics and focus specifically on the critical points in AI applications, depicted as the governance phase in figure 2.

As each military mission has unique requirements for the AI system, the third phase considers mission-specific adaptation. Starting with the definition of the mission goal in the system's (common) language and in accordance with the system's inherent engagement rules, this phase may be described as »parameter-setting«. The AI applications have been fully trained and checked in the previous phases, so this adaptation phase refers to parameter settings for the specific mission goals, available resources and environmental conditions, similar to how it is done in briefings for the military personnel going on this mission. This adaptation has to be done by military personnel.

The last phase is the deployment and use of the resulting AI system. Here two military roles are the main stakeholders: the operator of the AI weapons system and a supervisory authority dealing with the planning and analyses of operations.
The system's AI components act as an assistance for the weapons operator with the possibility of an explicit explanation facility to explain why a suggestion has been made by the system or not. The architecture of the four phases will be described in more detail below.

Development phase

Figure 3 sketches the components of an AI-based weapons system in the development phase. The core element is the AI storage system. It contains the foundation for the **»moral behaviour« and explainability of the AI weapons system**. On the one hand, this module defines a common language for all actors to specify unambiguous engagement rules as well as the system's capabilities in a machine and human interpretable form. A language always consists of a grammar and a vocabulary.

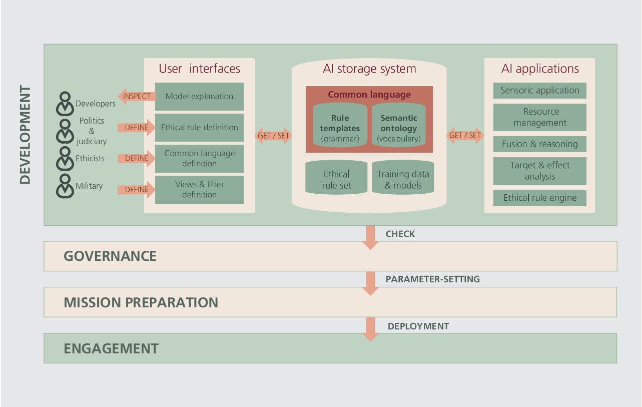

Figure 3: Core elements in the development phase of an AI-based weapons system.

As AI systems will make extended use of machine-learning algorithms,the used training data has to be centrally stored in addition to the models for the classical AI approaches. The AI algorithms of the AI weapons system are described in the AI application module. Four groups of AI components and an ethical rule engine are defined in the architecture; AI-based sensor applications for detection, classification and analysis of object and behaviour of interest make it possible to detect external circumstances. AI-based resource management allows the use of the weapons system's resources to be optimised. Fusion and reasoning modules will deliver an enhanced situational understanding. Target and effect analysis has to match the engagement rules with the weapon's capabilities. These four capabilities rely on the ethical rule engine to adjust their behaviour to the ethical rule set. This guarantees that the ethical guidelines are used to implement and execute the AI components of the AI system.

To define and examine the content of the AI storage system, user interfaces are necessary. Different views for each group of actors must be available to enter and examine ethical rules, to define the common language, and to get explanations for system suggestions using XAI (explainable AI) technology.

Adequate processes, methods and technologies to ensure trustworthiness must be established in the development phase, constituting a **trustworthiness engineering framework**. Such a framework could be inspired by existing engineering approaches and guidelines for high-integrity systems (e.g. safety engineering and safety standards) and it would cut across the complete development phase. The end point of the trust engineering activities is a trust assurance case, i.e. a comprehensive argumentation of the trustworthiness of the overall system in its context. The argumentation comprises all relevant requirements of trustworthiness as well as sufficient evidence to prove that all these requirements are met by the system.

Governance phase

The next step after the development is the governance phase (figure 4). The AI-based weapons system under consideration is fully developed at this point, meaning that all machine-learning parts are fully trained and the models and rules are stabilised. The resulting algorithms are static but may contain parameters for specific adjustments; this will be described in the third phase in detail. **This phase serves to check the compliance with the legal and ethical standards**. A core item for this check is the trust assurance case, which provides the overall trustworthiness argumentation bolstered by evidence and interlinked with all relevant development artefacts. Evidence will include test results, but there will also be independent tests conducted by the qualification bodies. The modules to be independently tested are the AI storage systems content (rules, test and training data) and the AI applications with the rule engine (figure 4).

Our recommendation is that modules have to be qualified as well as the system as a whole. Rules have to be checked for correctness, training data for fair and realistic balance, models for correctness and appropriateness. The rule engine has to be tested for its handling of priorities and discrepancies.

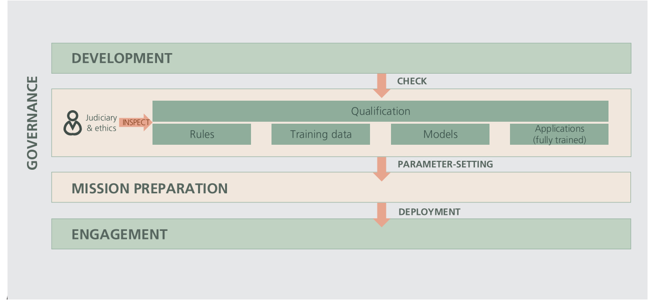

Figure 4: Elements of an AI-based weapons system to be qualified.

Mission preparation phase

Each military mission has unique requirements for the operations, e.g. peacekeeping, humanitarian or combat mission. Missions take place in different parts of the world, based on different rules of engagement and different laws. Therefore, the developed AI system needs to be adaptable to specific mission goals. In the mission preparation phase (figure 5), the mission goal has to be specified using the common language.

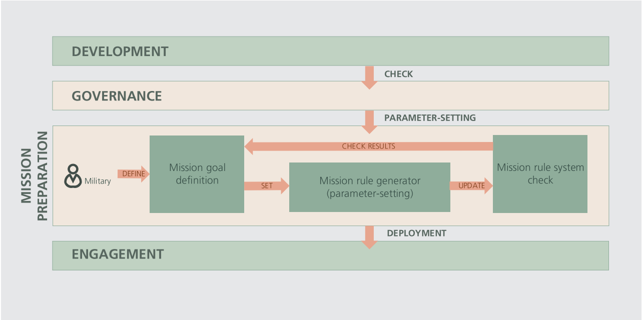

Figure 5: Mission preparation phase.

Parameter-setting for the algorithms (e.g. environmental conditions) and adjustment of the rule system (e.g. task-specific rules of engagement, different priorities) will prepare the system for engagement. A mission rule generator will set the parameters for the AI system according to the mission goal. This has to be checked to obviate errors and discrepancies between mission goal, rules and usable allowed resources to achieve the mission goal.

The overall parameterisation pattern must be adequately reflected in the trust assurance case to ensure that trustworthiness is guaranteed in any conceivable mission context.

Engagement phase

This phase (figure 6) covers the use of the AI system in a specific military mission. The mission-specific adaptation (parameter-setting) and thereby the mission goal and the rule engine are encapsulated in the mission-specific rule dataset. Building upon this, four modules assist the user. The sensor applications unburden the user from handling raw data. The quality of detection, classification and tracking results will be increased by using AI-based data analysis, fusion and reasoning algorithms. Intelligent resource management supports the user by optimising the potential weapons resources. Most ambitious is assistance for target and effect analysis.

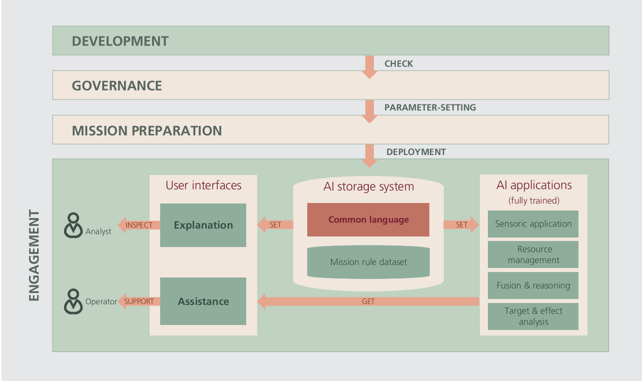

Figure 6: Engagement phase.

In order to follow the engagement rules, e.g. not to use a 50 mm weapon on human targets, or minimising collateral damage, the mission goal and available resources must be matched. In order to do so, real-time simulation calculating potential effects in the actual environment has to be used. Besides the user interface for the operator that just enables or disables functions, there has to be a special user interface supporting the inspection of the AI proposal, asking why the system made a proposal to apply a weapon with a special parametrisation or not. Explainable AI is the key for this. On the other hand, the user must be properly trained to interact with the system.

There will likely (depending on the actual application) be corresponding assumptions in the trustworthiness assurance case and it is then key that the assumptions always hold.

5. RESEARCH NEEDED

AI-based weapons systems have to follow the **governing laws, pertaining rules of engagement and ethical standards** and need to be **strictly qualified concerning their compliance** with these. Autonomous actions against human targets are at least ethically questionable and their fielding should be scrutinised carefully in every single case. Human operators and decision makers need to be fully responsible for every action taken with every weapons system independent of the integration of AI in the system. However, due to complexity, data amount and time conditions there are situations where machines have to react automatically. Therefore, the process of developing and deploying AI-based weapons must be strictly controlled and the trustworthiness of the overall system must be ensured. The previous chapters described a corresponding approach and outlined different phases and basic elements. In order to operationalise the envisioned approach, intensive research is required in several fields.

In the following, we present a brief description of the most urgent topics for interdisciplinary research.

- One key requirement is to establish a common language based on existing technologies. This will enable a sufficiently formal **specification of an AI-based system** (e.g. with respect to its requirements and use-cases) as well as provide a definition of all relevant terms.
- **Established policies, laws, regulations and norms shall also be specified based on this language, similar to the ethical goal** function described in [14] or a policy enforcement approach, similar to data usage control, described in [15]. In doing so, a conformity assessment of general system properties and of mission goals becomes possible.
- Moreover, it shall be investigated how exactly a **trust assurance case can be engineered for systems with AI components and how it is seamlessly linked to the ethical and legal boundaries specified in the common language**. An assurance case process might be established to give corresponding guidance to engineers.
- A key aspect regarding **»technical trustworthiness«** (safety, reliability and robustness in particular) is the utilisation of techniques from the field of **explainable AI** (as well as other types of analyses) for analysing, understanding and hardening the AI components and for creating evidence to bolster the claims made in the assurance case. A further use case for **explainable AI** is to enable systems to **explain their reasoning** to their operators in the field. It has to be investigated in which way corresponding analyses might be used for which cases (i.e. types of AI) and how the results are best presented to the operator (e.g. based on the common language).
- That apart, **security aspects** of the envisioned systems need to be investigated. In particular, adversarial attacks with respect to the AI component must be taken into account and means must be investigated to assess the susceptibility of a given component and to **harden it against manipulation and misuse**. A further security aspect concerns the **upgradeability and evolution** of systems, as it must be ensured that the trustworthiness of the system is never compromised.

Overall, there is an urgent **need for continuously dealing** with possible scenarios and use cases corresponding to the increasing potential of AI-based weapons systems.

6. SUMMARY

Weapons systems containing artificial intelligence will change the nature of armed conflicts on a fundamental level. The first of these new weapons systems are being fielded and more are in development. The question if AI will enter warfare has been answered. The open question is how this procedure will be shaped by humans – the ones directly responsible but also society as a whole. The central and important aspects of this implementation procedure are the complex interplay of **trustworthiness, controllability, predictability, explainability and the question of both responsibility and accountability**. This interplay will enable the acceptability of AI-based weapons systems.

To humans used to understanding conventional machines, AI methods are by their nature not transparent. In order to trust an AI system and make it transparent, means to **explain and predict the behaviour of AI systems in an understandable and coherent way are a necessary requirement**. One of the goals of current AI research is to identify approaches for these open questions, among them the promising concept of a common language spoken by humans and AI systems. This is also understood to be a feasible approach to exert control over AI systems and set **hard boundaries and precise goals**.

Only if the **required level** of trustworthiness is achieved can the question of responsibility be addressed by defining the role of AI systems in the complex military apparatus. This will be accompanied by the necessary redefinition of the human role in this context: what minimum level of control must a human have and what level of automatisation or even autonomous action can we transfer to machines while still being ethically and legally **responsible for the actions?**

While there are currently no answers to these questions, Fraunhofer VVS hopes that this position paper will provide impetus for possible approaches to finding them.

7. REFERENCES

[1] A. Jobin, M. Ienca and E. Vayena: Artificial Intelligence: the global landscape of ethics guidelines, Nature Machine Intelligence, Vol. 1, No. 9, pp. 389–399, 2019.

[2] W. Koch: Zur Ethik der wehrtechnischen Digitalisierung. Informations- und ingenieurwissenschaftliche Aspekte. In: M. Rogg et al. (Ed.): Ethische Herausforderungen digitalen Wandels in bewaffneten Konflikten. GIDS, Hamburg 2020 (E-Book), Open Access, https://gids-hamburg.de/ethische-herausforderungen-digitalen-wandels-in-bewaffneten-konflikten, pp. 17–54.

[3] BMVg: Die Konzeption der Bundeswehr, Bundesministerium für Verteidigung, 2018.

[4] M. J. Wooldridge: An introduction to multiagent systems, 2nd ed. John Wiley & Sons, 2009, p. 461.

[5] R. Onken and A. Schulte: System-ergonomic design of cognitive automation, Vol. 235, Heidelberg, Springer, 2010, p. 383.

[6] BMVg: Erster Bericht zur Digitalen Transformation des Geschäftsbereichs des Bundesministeriums der Verteidigung, Bundesministeriums der Verteidigung, Berlin, 2019.

[7] Oberstleutnant T. Doll and Hauptmann T. Schiller: Künstliche Intelligenz in den Landstreitkräften – Ein Positionspapier des Amtes für Heeresentwicklung, Amt für Heeresentwicklung, 2nd ed., 2019.

[8] J. McCarthy, M. L. Minsky, N. Rochester and C. E. Shannon: A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, AI Magazine, No. 27, 1955.

[9] W. Graf von Baudissin: Soldat für den Frieden. Entwurf für eine zeitgemäße Bundeswehr. Beiträge 1951–1969, Piper Verlag GmbH, 1982.

[10] F. Sauer: Kennzeichen des Krieges im 21. Jahrhundert, Außerschulische Bildung, Arbeitskreis deutscher Bildungsstätten e.V., 2017, pp. 4–10.

[11] General J. R. Allen, U.S. Marine Corps (Retired) and A. Husain: On Hyperwar, U.S. Naval Institute, Proceedings, No. 143, July 2017.

[12] Oberstleutnant T. Doll, U. Beyer and Hauptmann T. Schiller: Hyperwar – Neue Herausforderungen für die Heeresentwicklung, Europäische Sicherheit & Technik (ES&T), 9/2019.

[13] High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI, European Commission, 8 April 2019.

[14] P. Elands, A. Huizing, L. Kester, S. Oggero and M. Peeters: Governing ethical and effective behaviour of intelligent systems, TNO Defense, Safety and Security, The Netherlands, White Paper, Oct. 2018

[15] P. Birnstill: Privacy-Respecting Smart Video Surveillance Based on Usage Control Enforcement, Dissertation, Karlsruher Schriften zur Anthropomatik, Vol. 25, Karlsruher Institut für Technologie (KIT), 2016.

Imprint

Editors
J. Beyerer, Fraunhofer IOSB, Chairman of Fraunhofer VVS
P. Martini, Fraunhofer FKIE, Deputy Chairman of Fraunhofer VVS

Text
R. Schönbein, Fraunhofer IOSB,
A. Hoffmann, Fraunhofer IOSB,
W. Koch, Fraunhofer FKIE,
S. Moser, Fraunhofer EMI,
V. Heusinger, Fraunhofer EMI,
K. Pervölz, Fraunhofer IAIS,
D. Schneider, Fraunhofer IESE,
S. Brüggenwirth, Fraunhofer FHR,
W. Eckl, Fraunhofer ICT,
O. Gabel, Fraunhofer INT

Contact
Caroline Schweitzer, Fraunhofer IOSB
Gutleuthausstr. 1, 76275 Ettlingen, Germany
caroline.schweitzer@iosb.fraunhofer.de

Images
Cover, The Future Combat Air System, © Fraunhofer FKIE
Page 2, © Fraunhofer IOSB, indigo

Layout
B. Euscher, Fraunhofer IOSB

www.vvs.fraunhofer.de