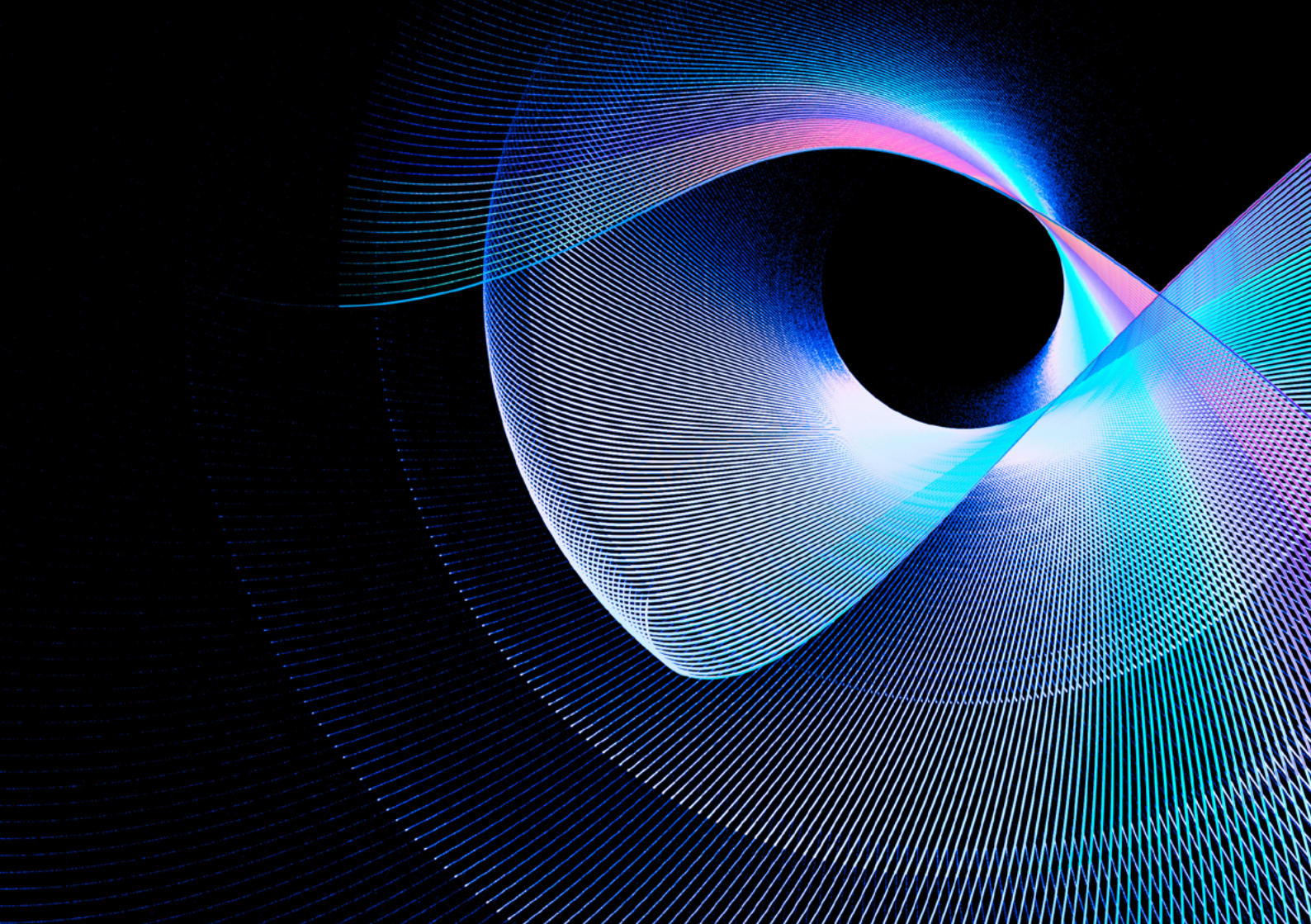
EUROPEAN DEFENCE AGENCY

WHITE PAPER

TRUSTWORTHINESS FOR AI IN DEFENCE

Developing Responsible, Ethical, and Trustworthy AI Systems for European Defence



Trustworthiness for AI in Defence (TAID)

-

White Paper

Developing Responsible, Ethical, and Trustworthy AI Systems for European Defence

Publication date: 09/05/2025

Document version: 1.0

Table of Contents

List of Figures

List of Tables

Document Information

Classification

This document is Unclassified.

Conditions of release

This document is delivered by the European Defence Agency to the EU Defence Community.

Version history

HISTORY OF CHANGES				
Version	**Date**	**Reason of Changes**	**Reviewed By**	**Approved By**
0.1	11/10/24	First Draft	TAID WG Members	
1	09/12/24	Submission for approval by MoDs	EU MoDs	EU MoDs
1.1	31/03/25	Submission to EDA for publication approval	EDA	

Document Authors

Name	**Entity**
Isidoros Monogioudis	EDA
Luis Javier Costa Giraldo	INDRA
Dr. Andreas De Jonge	DE JONGE GmbH
Bruno Carron	AIRBUS DE
Marcilli Gianluca	IT MOD
Alison M. Kay	TCD IE
Daniele Bet	IT MOD
Janaina Ribas De Amaral	AIRBUS DE
Gabriel Pedroza	ANSYS
Fateh Kaakai	THALES
Michel Barreteau	THALES
Andreas Tollkühn	MBDA DE
Luciana Morogan	RO MOD
Liisa Janssens	TU DELFT
Yvonne Hofstetter	21 STRATEGIES
Chrystèle Johnson	MDBA FR

Document Contributors

Name	Entity
Monika Venckauskaite	BPTI
Alessio Cavallin	LEONARDO
Fabio Magosso	LEONARDO
Javier Ferrero Micó	GMV
Arnold Akkermann	DLR DE
Dr. Markus Hosbach	IABG
Antonio Monzón-Diaz	AIRBUS DE
Joseph Machrouh	THALES
Simon Bensberg	RHEINMETALL
Dr. Frank Beer	INFODAS
Alexis De Cacqueray	AIRBUS DE
Andromachi Papagianni	CERTH
Christophe Guettier	SAFRAN ELECTRONICS & DEFENSE
Papantoniou Vassilios	HTR
Ruben Post	TNO NL
Tuulia Timonen	CGI

Contact

Any comments should be addressed to:

European Defence Agency

Research, Technology and Innovation Directorate
Rue des Drapiers 17-23
B-1050 Brussels
Belgium
Email: RTI@eda.europa.eu
Web: www.eda.europa.eu

1. Introduction

The purpose of this document is to collect, present and describe the aspects of Trustworthiness for AI in Defence in a 'food for thought' approach reflecting the combined view of AI experts and stakeholders from Defence Industry, Academia and Ministries of Defence. This effort is performed in the context of the European Defence Agency's (EDA) Action Plan on Artificial Intelligence for Defence and tries to address the topics of trusted AI and verification, validation and certification requirements analysis. The topics covered and analysed in this document will provide the appropriate knowledge of the current global status considering the AI regulations, standards and frameworks for AI trustworthiness and will also recommend the follow-up activities that will further assist the EU Members States and Defence Industry to better prepare, plan and develop the future AI systems aligned with the identified expectations.

The target audience is EU Member States especially the MODs that will further evaluate the whitepaper's recommendations and may use it as a reference point for future related AI research activities. Target audience will be expanded also to Defence Industry to highlight the key aspects and requirements for developing AI systems for military use considering all regulations, standards and methodologies assisting them in better planning and design to deliver trusted AI services and applications.

The report is structured based on the identified tasks as follows:

- AI definition and Taxonomy
- Identification of stakeholders of AI in defence
- Legal perspective for AI use cases and scenarios
- Standards and Regulations
- Testing & Evaluation, Validation and Verification standards, tools and methodologies
- Human Factors in Defence
- Ethical AI considerations
- Impact analysis of AI in Defence
- Military Use cases and scenarios for Trustworthiness of AI
- Recommendations and next steps

This whitepaper is the starting point and an initial reference source for future related AI research activities and project proposals that essentially will provide to the community the adequate information and knowledge of how to plan, develop, acquire, test and use defence AI Systems.

The whitepaper is also affected by the recently approved EU AI Act [1] which is a regulation that is expected to have a significant impact on the future design and use of AI technology across EU.

As part of the latest updates regarding the AI Act's progress [1], the Commission issued a standardisation request, tasking CEN and CENELEC with developing European standards by 30 April 2025. These standards aim to ensure AI systems in the EU market are safe, uphold fundamental rights, and encourage innovation.

Either way the recommendations of this document, without raising any commitment or obligation from the Member States, are expected to feed partially the proposed AI Action Plan update (v.2.0) that will further address in more detail the topics raised both in this paper but also any other identified topic that will be introduced and accepted over the action plan's development process.

It must be noted that this whitepaper is mainly an outcome of a work done in a volunteer basis by the Trustworthiness for AI in Defence Working Group with multidisciplinary expert members from Industry and Academia across Europe but also members from some EU MODs.

2. Scope: AI Definition and Taxonomy

2.1. AI–based system Definition

During the past years there have been many attempts to formulate a clear definition of Artificial Intelligence (AI). These attempts were accompanied by fundamental discussions about human intelligence in general and machine intelligence. One major outcome of this process was to differentiate between the research on (artificial) intelligence and engineered systems leveraging the results of this research. This whitepaper will build on these findings and will focus on engineered systems, meaning that every time the term AI is mentioned it should be understood as an engineered system and not as the research discipline of AI in general. Therefore, inspiration is taken from the European civil sector and the definition for Artificial Intelligence from the European AI Act is used:[1]

The definition of an AI-based system in Article 3(1) has been modified to align it more closely with the work of international organizations working on artificial intelligence, notably the Organisation for Economic Co-operation and Development (OECD). Moreover, the corresponding Recital 6 details further the key characteristics of the definition and clarifies that the definition is not intended to cover simpler traditional software systems or programming approaches, which are based on the rules defined solely by natural persons to automatically execute operations. Additionally, the Commission has been tasked to develop guidelines on the application of the definition of an AI-based system.[2]

The definition provided by OECD is as follows:

"*An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment*." See [2] for further detail.

To ensure effective governance and compliance in these AI systems, organizations should implement data dictionaries and business glossaries. These resources provide standardized definitions and promote a common understanding among stakeholders, ensuring clarity and consistency in data interpretation as well as a shared understanding of key terms. Regular maintenance of these resources is crucial to adapt to changes in data usage and regulatory requirements, thereby supporting ongoing compliance and effective governance. Implementing a robust change management process is essential for managing updates to these tools, ensuring that all stakeholders are informed of changes, and that the definitions remain accurate and relevant.

Reflecting the special needs, use cases and scenarios of the defence sector, this definition needs to be expanded to cover the topics that are specific to military use of AI. Specific characteristics to consider include operational concepts and doctrines, values, and the adversarial environment in which defence applications are used.

[1] AI Act 12th July 2024 version

[2] This Regulation shall not apply to areas outside the scope of EU law and in any event shall not affect the competences of the Member States concerning national security, regardless of the type of entity entrusted by the Member States to carry out the tasks in relation to those competences. This Regulation shall not apply to AI systems if and insofar placed on the market, put into service, or used with or without modification of such systems exclusively for military, defence, or national security purposes, regardless of the type of entity carrying out those activities. This Regulation shall not apply to AI systems which are not placed on the market or put into service in the Union, where the output is used in the Union exclusively for military, defence, or national security purposes, regardless of the type of entity carrying out those activities.

2.2. Taxonomy for Defence

Up to now, the EDA's OSRA Defence Technology Taxonomy v2.0 [3] (OSRA Defence Technology Taxonomy) with a dedicated chapter for Specialized Technology Taxonomies – AI, has included little content about trustworthy AI, reflecting a gap in its coverage of this critical aspect. In the civil sector, standards for AI primarily come from organizations such as ISO and the IEEE. However, in the defence context, it is crucial to consider additional aspects specific to the operational and military domains, where AI systems must not only adhere to these standards but also meet stringent requirements for security, resilience, and performance under adversarial conditions.

The following paragraph highlights the aspects of AI in these domains and should serve as a starting point for a defence-specific high-level taxonomy to be included in the OSRA taxonomy.

2.2.1. Aspects of AI in the Technical Domain

The technical aspects of Artificial Intelligence (AI) are shaped by standards from ISO and the IEEE, which ensure AI technologies are safe, reliable, and ethically aligned. ISO standards like ISO/IEC 22989 [4], ISO/IEC 23053 [5], and ISO/IEC 24029 [6] focus on framework, data quality, and system robustness, respectively, promoting consistency and reliability in AI systems. IEEE's 7000 series addresses ethical and technical aspects, with IEEE 7010 [7], 7012 [8], and 7001 [9] focusing on human well-being, algorithmic transparency, and ethical considerations such as fairness and privacy.

Adherence to these standards involves managing high-quality, unbiased data, rigorous model testing, ensuring interoperability, and embedding ethical considerations into AI design. This structured approach fosters global collaboration and trust, enabling the development of AI systems that are both technically robust and socially responsible.

2.2.2. Aspects of AI within Operational Domains

The operational domain refers to the specific environment or context in which a system, process, or activity operates. In fields such as military or technology, the operational domain defines the scope and conditions under which operations are conducted.

In military terminology, the operational domain encompasses the geographic area, airspace, maritime space, and cyberspace where military forces conduct their missions and activities. It includes factors such as terrain, weather, enemy forces, and civilian populations that influence military operations.

In technology, the operational domain can vary depending on the specific application or system. For instance, in cybersecurity, the operational domain may refer to the network infrastructure, software applications, and data assets that are protected and monitored for security threats.

For that reason, trustworthiness can be seen from different dimensions i.e. military, technical, and legal domain (wherein human factors play a role in the human-machine cooperation). The following sections describe these domains in detail.

2.2.3. Aspects of AI in the Military Domain

In the military context, trustworthiness is paramount due to the critical nature of military missions and the potential consequences of AI system failures or misuse. Trustworthy AI systems in the military domain are designed to inspire confidence among commanders, soldiers, and policymakers by ensuring that AI technologies operate predictably, securely, and

ethically in complex and dynamic environments. Therefore, the main aspect in the military domain is criticality which can be addressed as follows:

Criticality: different methods exist to assess criticality according to the application domain. In general, the risks are calculated as a trade-off between probability of feared events and the severity of their impacts (other parameters can be considered). To assess the criticality and acceptability of residual risks, several classifications exist among them: Safety Integrity Levels (SIL) IEC-61508 [10], EN-50129 [11], Automotive Safety Integrity Levels (ASIL) ISO-26262 [12] and Development Assurance Levels (DAL) ARP-4754b [13].

For military application, the following considerations should be taken into account:

- Hierarchical organization encompassing collaborative AI customers and AI operators.
- Respect an operational doctrine involving rules of engagement and war-fighter collective values: responsible usage, commanding trust, subsidiarity principles, mutual support, endurance, and resilience.

AI customer and AI operator activities can be structured by the OODA loop. In current operations most decisions are made by humans (exception: Iron Dome, vessel proximal defence). We expect that in the future levels of autonomy will increase as well as we will see new types of autonomy, human in the loop, human on the loop, human out of the loop.

2.3. Stakeholder of AI for Defence

2.3.1. Standards covering AI Stakeholders

Since AI is a transdisciplinary field, finding clear definitions for stakeholders and their roles is challenging. In the context of Data and AI Governance, which encompasses Governance, Risk, and Compliance (GRC) principles, stakeholder identification and role definition become crucial for establishing accountability and ensuring responsible AI practices. By analogy with the civil domain the following documents can serve as a starting point for further refinement of the description of stakeholders and their roles in AI applications.

- ISO/IEC DIS 12792 (under development) [14]
- ISO/IEC 22989:2022 (published) [4]
- ISO/IEC TR 24028:2020 (published) [15]
- IEEE 7000-2021 (published) [16]
- ISO/IEC 42001:2023 (published) [17]

Especially the ISO/IEC 22989:2022 [4] delivers an appropriate overview over stakeholders in AI and their (sub) roles for the civil sector. Coming from this top-down categorical description covering generic terms and roles, the provided stakeholder definition proposed in this document can be interpreted as a road map to define AI-stakeholders and their roles with respect to the different time frames (i.e. strategical, operational, and tactical) and use cases resulting in a scenario-based approach.

2.4. Terms and Roles

As explained before, the scenario of an AI-application is created by the operational level and the use case of the application. Diagram presented in Figure 1, derived from ISO/IEC 22989:2022 [4], reflect the proposed stakeholders and their roles for defence applications. The proposed operational levels feed into the scenario leading to the extraction of AI-stakeholders. It should be pointed out that unlike some civil usages, military usages of AI

usually require a formal transfer of risk and ownership from the supplier to the user, for the reason that the usage of force is a prerogative of the State. Green fields in Figure 1 relate to terms which are directly taken form ISO/IEC 22989:2022 [4]. The explanations for terms therein can be found in the Appendixes. The red fields in Figure 1 show the proposed additions for defence applications.

2.4.1. Strategic, Operational and Tactical Level

Reflecting the requirements of the defence sector, different levels need to be included into stakeholder definitions influencing the role and activities of different stakeholders. These levels are usually divided into strategical, operational, and tactical level.

On the **strategic level**, stakeholders engage in activities focused on long-term planning, goal-setting, and strategic decision-making. This includes developing overarching strategies, identifying opportunities, and aligning organizational objectives with trends and industry developments. Stakeholders at this level may also be involved in forecasting future challenges and opportunities, as well as establishing partnerships and alliances to support long-term objectives.

On the **operational level**, stakeholders implement strategies outlined in the strategic plan through specific projects, initiatives, and campaigns. This involves allocating resources, defining key performance indicators (KPIs), and monitoring progress towards strategic goals. Operational activities may include launching new products or services, optimizing operational processes, and adjusting strategies based on feedback.

On **the tactical level**, stakeholders focus on day-to-day activities necessary for the execution of tactical plans. This includes managing daily operations, coordinating tasks and activities, and resolving any issues or challenges that arise in real-time. This also includes the actors and operators of AI based systems on the battlefield during a conflict or a crisis management situation.

Together with the use case in which context the AI application should be brought into action these levels result into a scenario from which the different stakeholders and their roles can be derived. Section 8.4**Error! Reference source not found.** shows some examples for this a pproach.

Considering the information presented before, this approach can be used to identify stakeholders based on the specific scenario in which an AI-based system is implemented. The scenario is determined by the use case and the level of the AI system. Stakeholders are identified through a top-down method, starting with the generic Term for an AI stakeholder, the level and general categories such as AI provider, producer, customer, partner, subject, and relevant authorities. Green fields represent stakeholders and roles already acknowledged in ISO/IEC standards, while red fields highlight proposed additions specific to the defence sector. This method aims to ensure a comprehensive understanding of all relevant stakeholders involved in deploying AI systems in defence contexts.

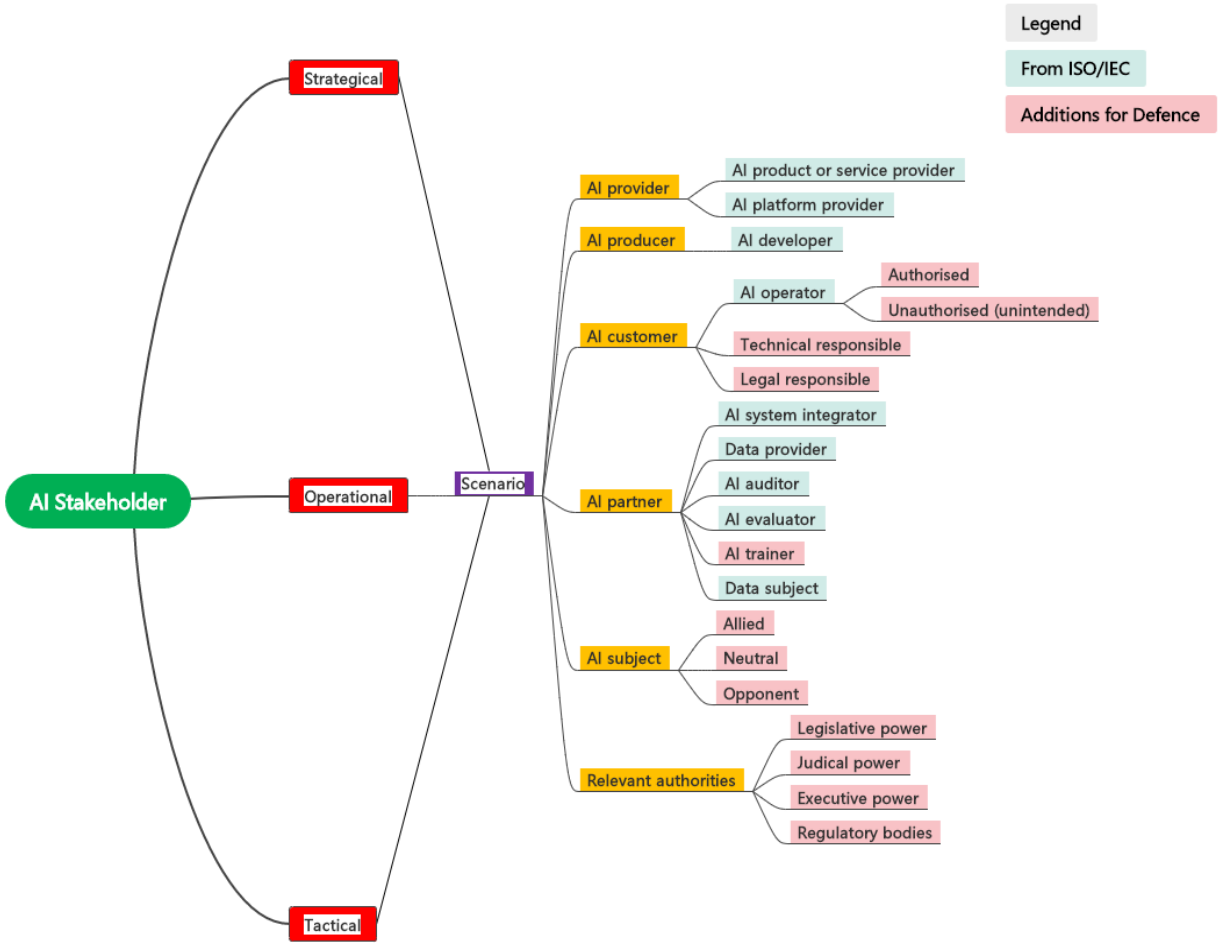


Figure 1 - Overview of AI stakeholders and roles

Stakeholders are derived top-down, starting with the generic AI stakeholder, the level and general categories like AI provider, producer, customer, partner, subject, and relevant authorities. Green fields show stakeholders from ISO/IEC standards, while red fields indicate proposed defence-specific additions.

2.4.2. Stakeholder Roles

The following list contains a collection of the terms with their explanations shown in the figure above.

- **AI Provider**
  - General: An AI provider is an organization or entity that provides products or services that uses one or more AI systems. AI providers encompass AI platform providers and AI product or service providers.
    - AI product or service provider: An AI service or product provider is an organization or entity that provides AI services or products either directly usable by an AI customer or user, or to be integrated into a system using AI along with non-AI components.
    - AI platform provider: An AI platform provider is an organization or entity that provides services that enable other stakeholders to produce AI services or products.

- **AI Producer**
  - General: An AI producer is an organization or entity that designs, develops, tests and deploys products or services that use one or more AI system.
    - AI developer: An AI developer is an organization or entity that is concerned with the development of AI services and products.
- **AI Customer**
  - General: An AI Customer is an institution/ organization or entity that uses AI products or services.
    - AI Operator: An AI Operator is an individual or entity responsible for overseeing the functioning of an AI system. Their duties include managing the system's operations, ensuring its performance aligns with intended purposes, and intervening when necessary to prevent adverse outcomes or to enhance system efficacy. The AI Operator also plays a crucial role in monitoring ethical concerns, managing security risks, and ensuring compliance with relevant regulations and standards. [17]
      - Authorized AI Operator: An authorized AI operator refers to an individual or entity that has been granted permission or authorization to access certain resources, systems, or information within an organization. This permission is typically granted by an administrator or manager who oversees the relevant assets.
      - Unauthorized (unintended) AI Operator: An unauthorized (unintended) AI operator refers to an individual or entity that does not have permission or authorization to access specific resources, systems, or information within an organization. Unauthorized users may attempt to gain access through various means, such as hacking, social engineering, or exploiting vulnerabilities in security systems. Their actions are typically in violation of organizational policies and legal regulations governing data protection and privacy.
    - Technical Responsible: A technical responsible customer is an organization or entity that takes charge of technical implementation and performance, ensures system integration, addresses technical issues, and collaborates with providers to meet operational needs, while ensuring compliance with security and regulatory standards.
    - Legal Responsible: A legal responsible customer is an organization or entity accountable for ensuring compliance with applicable legal and regulatory frameworks. This includes overseeing data privacy, intellectual property rights, contractual obligations, and ensuring that the deployment of AI systems adheres to laws regarding fairness, transparency, and accountability. The legal responsible customer works closely with legal advisors to mitigate risks, address liability concerns, and ensure that AI operations remain compliant with evolving laws and ethical standards.
- **AI Partner**
  - General: An AI partner is an organization or entity that provides services in the context of AI. AI partners can perform technical development of AI products or services, conduct testing and validation of AI products and services, audit AI usage, evaluate AI products or services and perform other tasks. Examples of AI partner types are discussed in the following subclauses.
    - AI system integrator: An AI system integrator is an organization or entity that is concerned with the integration of AI components into larger systems, potentially also including non-AI components.

- Data provider: A data provider is an organization or entity that is concerned providing data used by AI products or services.
- AI auditor: An AI auditor is an organization or entity that is concerned with the audit of organizations producing, providing, or using AI systems, to assess conformance to standards, policies, or legal requirements.
- AI evaluator: An AI evaluator is an organization or entity that evaluates the performance of one or more AI systems
- AI trainer: An AI trainer is a person, or an entity specialized in the training of users to apply specific AI systems.

- **AI Subject**:
  - General: An AI subject is an organization or entity that is impacted by an AI system, service, or product.
    - Data subject: A data subject is an organization or entity that is affected by AI systems with following aspects: Subject of training data: where data pertaining to an organization or human is used in training an AI system, there can be implications for security and privacy, for the latter particularly where that subject is an individual human.
    - Allied: Allied AI subjects focus on developing AI algorithms that support the goals or operations of another AI system or entity, fostering synergy and cooperation to achieve common objectives efficiently.
    - Neutral: Neutral AI subjects focus on advancing AI technologies for general purposes across various domains, without aligning with or opposing any specific entity or goal.
    - Opponent: Opponent AI subjects focus on designing AI systems to act against specific targets, often with hostile intent. They focus on developing algorithms and strategies to disrupt, infiltrate, or undermine adversaries' operations or security.
- **Relevant Authorities**
  - General: Relevant authorities are organizations or entities that can have an impact on an AI system, service, or product. Separation of powers (the separation of the legislative, judicial, and executive power) is important to adhere to and checks & balances need to be strengthened with additional requirements. These requirements can enrich existing mechanisms in order to foster the delicate balancing act of navigating in a responsible way difficult processes such as procurement and deployment. Furthermore, shaping public-private partnerships - in the scope of AI- need to be investigated how trustworthy AI can be established.
    - Legislative power: Government that has the authority to set laws within a national, European, and non-EU level that impact processes such as the procurement, development and deployment of AI systems, services, or products.
    - Judicial power: Judges of international, European, and national courts who need to be able to judge if an AI system, service, or product adheres to laws, rules, and regulations.
    - Executive power: This is enforcement of the law, of which the military and law enforcement play a key role.
    - Regulatory bodies:
      - Non-Governmental
      - Standardization Organizations
      - EDA
      - NATO

3. Legal perspective for AI use cases and scenarios

European norms and values shape roles and mandates to protect society from internal and external threats. Trust and the legal system are complementary but cannot replace each other. Trust is based on the legal system, which only functions when there is trust in its operational aspects, such as checks and balances founded on legitimacy and the separation of powers. In military operations, a commander's trust in AI systems is crucial. For instance, how can a commander ensure compliance with the rules of engagement when using AI for decision support? Furthermore, military operations value secrecy, which often conflicts with the need for transparency. How can we reconcile these conflicting requirements?

The characteristics of a constitutional society include the separation of powers and checks and balances. The Rule of Law shapes roles and mandates to protect society from internal and external threats. The legislative, executive, and judicial branches must operate within the Rule of Law, with the executive branch, including law enforcement and the military, playing a key role in society's protection. Their methods must be adequate, efficient, and trustworthy. [18]

It is essential to recognize that the Rule of Law is not merely 'positive law,' which consists of compliance rules and regulations. The Rule of Law involves principles that apply directly to real-life use cases and scenarios, derived from various sources such as case law, legal doctrine, interpretation methods, positive law, draft regulations, and legal theory. Existing mechanisms of the Rule of Law can be observed in processes like new legislation or policy redefinition within legal boundaries. [18]

Given the complexities involved, especially in military operations where secrecy and transparency often conflict, a holistic approach might be more fruitful. This means considering a broader perspective that goes beyond immediate tactical concerns. We must design scenarios that address operational levels comprehensively and cater to each specific use case. This approach ensures that the integration of AI in military operations is done thoughtfully, maintaining trust, compliance with the Rule of Law, and operational effectiveness.

By incorporating this broader perspective, leads to the following advantages:

- Enhance decision-making: Design AI systems that provide decision support aligned with the rules of engagement and legal frameworks, ensuring commanders can trust these systems in high-stakes environments.
- Balance secrecy and transparency: Develop protocols that protect sensitive information while maintaining necessary transparency to ensure accountability and adherence to legal standards.
- Ensure adequacy and efficiency: Create scenarios that test AI systems for their adequacy and efficiency in various operational contexts, ensuring they meet the demands of real-life military applications.
- Foster trust and legitimacy: Implement AI solutions that enhance the trust of military personnel by demonstrating reliability, fairness, and compliance with ethical and legal norms.
- Adapt to dynamic threats: Design flexible scenarios that can adapt to evolving threats and operational conditions, ensuring that AI systems remain relevant and effective in different contexts.

By focusing on the bigger picture and tailoring scenarios to specific operational needs and use cases, AI can be effectively integrated into military operations, enhancing both strategic and tactical capabilities while upholding the fundamental principles of trust, legality, and ethical conduct.

3.1. Scenario Construction

In developing and deploying AI systems, a structured scenario-based approach [18] ensures alignment with practical scenarios, regulatory compliance, and operational effectiveness. The following step-by-step explanation clarifies this process.

3.1.1. Integration of Use Case and Scenario Design

The process begins by combining the specific use case, for example a Counter Unmanned Aircraft System or Unmanned Ground Vehicles which make use of a specific type of AI to enhance system capabilities, with its corresponding operational level. A scenario-based approach helps in seeking an integrated level of understanding. This integrated approach helps to inform and shape the design of a relevant scenario and to illustrate which capability can be filled with what type of AI to reach a military advantage. This scenario outlines the context and environment in which the AI application will be deployed, ensuring that it is realistic and aligned with the intended purpose.

3.1.2. Derivation of Regulatory Scope from Scenario

Once the scenario is established, it is analysed to determine the necessary scope for regulation and standardization. This step involves identifying the regulatory requirements and standards needed to ensure that the AI system operates safely, ethically, and in compliance with existing laws. Furthermore, an operational scenario helps to pinpoint what future aspects need to be regulated to manage risks and ensure proper governance and can also help in framing the more fundamental risks for maintaining principles such as the rule of law (art 2 Treaty on European Union [19]) and how to mitigate these risks throughout the value chain of partners (starting from procurement and R&D up to the deployment).

3.1.3. Capability Gap Analysis and Regulation Selection

Based on the identified regulatory scope, a capability gap analysis is conducted. This analysis assesses the current capabilities against the requirements highlighted by the scenario. It identifies gaps that need to be filled to meet the regulatory standards. Concurrently, appropriate regulations are selected to address these gaps. This step ensures that the AI system can meet the necessary technical, operational and regulatory requirements.

3.1.4. Problem Identification for AI Solution

The primary goal of these steps is to clearly define which specific problem the AI system aims to solve. By understanding the scenario and regulatory requirements, stakeholders can pinpoint the challenges that the AI solution needs to address. This helps in focusing the development efforts on solving a real and defined problem, ensuring that the AI provides practical value.

3.1.5. Assessment of Military Advantage

Finally, the potential military advantage provided by the AI system is evaluated in relation to the identified regulations. This involves analysing how the AI application can offer a strategic edge or operational benefit within a military context, while still adhering to the regulatory

constraints. This step ensures that the AI system not only complies with regulations but also enhances military capabilities effectively.

This structured approach ensures that the development and deployment of the AI system are well-grounded in realistic scenarios, comply with necessary regulations and provide tangible benefits, particularly in a military setting. The goal is to create a solution that is both effective and compliant with norms and values of the European Union and help in (re)shaping legal, standardization and ethical standards.

3.2. Requirement Identification

The responsible development and deployment of AI systems, especially in high-risk domains such as defence, require adherence to robust frameworks and standards. This requires the implementation of comprehensive Data and AI Governance frameworks, integrating key elements of Governance, Risk, and Compliance (GRC). Such frameworks establish a structured approach to managing data and AI systems, ensuring regulatory compliance, mitigating risks, and enforcing ethical standards throughout the AI lifecycle.

These frameworks ensure that AI technologies are not only technically sound but also align with ethical, legal, and societal values. By drawing on a variety of international and regional standards, and incorporating them into a robust GRC-based Data and AI Governance model, organizations can develop AI systems that are trustworthy, transparent, and accountable. This governance approach enables organizations to identify requirements, conduct thorough risk and impact assessments effectively, implement procedures and controls, and maintain compliance while fostering innovation.

This section shows how a requirement analysis can be performed once a scenario has been constructed following the approach proposed in this document, leveraging governance principles to ensure comprehensive coverage of technical, ethical, and regulatory considerations.

3.2.1. Frameworks for Responsible AI

The first point is to focus on identifying the key frameworks and standards that guide the responsible use of AI. Several international and regional standards can be considered here, depending on the scenario and jurisdiction. Key documents include:

- ISO/IEC 22989 (AI Concepts and Terminology) [4] and ISO/IEC 23053 [5] (Framework for AI Governance) which are designed to standardize AI development and implementation practices.
- OECD AI Principles, which emphasize AI's ethical and trustworthy deployment, aligning with values such as human-centeredness, transparency, and accountability. [20]
- NATO's revised AI Strategy 2024 [21], builds upon the foundation laid in 2021 and emphasizes the responsible development and use of AI technologies in defense and security. As such, it reaffirms the six Principles of Responsible Use (PRU) for AI [22]: Lawfulness, Responsibility and Accountability, Explainability and Traceability, Reliability, Governability, and Bias Mitigation. These principles, aligned with NATO's values, norms and international law, guide the Alliance's approach to AI adoption, ensuring ethical compliance and addressing potential risks.

  The PRUs for AI serve as a baseline for Allies in their use of AI for defence and security purposes, applying across the lifecycle of an AI capability without superseding existing national or international obligations. To operationalize these principles, NATO

established a Data and AI Review Board, which develops practical Responsible AI toolkits, guides implementation, and supports Allies in their effort. This initiative aims to accelerate AI integration within NATO while maintaining responsible practices.

Furthermore, NATO commits to collaborating with relevant international AI standards setting bodies to help foster military-civil standards coherence with regards to AI standards.

- EU's Ethics Guidelines for Trustworthy AI [23], which provide a set of principles such as respect for human autonomy, prevention of harm, fairness, and explainability. These are critical in the context of defence applications, given the emphasis on trust and human oversight in high-risk domains.
- Policy papers from bodies like the European Commission, emphasizing accountability and governance, such as the White Paper on AI or the AI Act.
- AIGA AI Governance Framework [24], provides a practice-oriented framework for implementing Responsible AI, adopting a systematic approach for AI Governance throughout the AI lifecycle. Key premises to consider are:
    - Tasks are mapped to the OECD's AI system lifecycle framework.
    - Supports compliance with upcoming EU AI Act.
    - Provides a template for decision-makers to address the key questions on the use of AI. Applying the framework to design and implement practices for using AI in a socially and ethically responsible manner.
    - Is value-agnostic facilitating the development and deployment of transparent, accountable, fair, and non-maleficent AI systems.
- ISO/IEC 42001 [17], offers valuable guidance addressing unique challenges AI poses, such as ethical, transparency and continuous learning considerations. It sets out a structured way to manage risks and opportunities associated with AI, balancing innovation with governance. It specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System (AIMS) within organizations, ensuring responsible development and use of AI systems. Both, ISO/IEC 42001 and NIST AI RMF [25] complement each other.
- NIST AI Risk Management Framework (NIST AI RMF 1.0) [25]: Complementing these broader standards, the NIST AI RMF provides a detailed and structured approach to managing the risks inherent in AI systems. The framework is divided into two parts:
    - Part 1 outlines the characteristics of trustworthy AI, which include being valid and reliable, safe, secure and resilient, transparent, explainable, privacy-enhanced, and fair. These attributes are aligned with the ethical and technical principles set forth by the EU and OECD, ensuring a cohesive approach across jurisdictions.
    - Part 2 presents four core functions – GOVERN, MAP, MEASURE, and MANAGE – which serve as practical steps for organizations to address AI risks at every stage of the AI lifecycle. The GOVERN function, in particular, aligns with the governance aspects of frameworks like ISO 23053 [5], ensuring that the organization's AI risk management processes are robust, while the MAP, MEASURE, and MANAGE functions provide more specific applications to individual AI systems, aiding in aligning operational AI systems with the broader principles of fairness, transparency, and accountability.

- ISO/IEC 23894:2023 – Artificial intelligence - Guidance on risk management [26]: This ISO standard, specifically focused on risk management for AI, further strengthens the ability of organizations to handle AI-related risks. It provides comprehensive guidance on identifying, assessing, and mitigating risks associated with AI systems, including risks related to biases, security vulnerabilities, and ethical concerns. ISO/IEC 23894:2023 aligns well with both NIST AI RMF 1.0 and the broader ethical frameworks from the EU and OECD by emphasizing a structured approach to identifying and mitigating risks at every stage of the AI lifecycle.
    - ISO/IEC 23894:2023 and NIST AI RMF 1.0 complement each other by addressing risk management from both an organizational and technical perspective. While NIST AI RMF 1.0 outlines specific risk management functions such as governance, mapping, measurement, and mitigation, ISO/IEC 23894:2023 provides additional best practices and guidelines on how to systematically address these risks, particularly in AI systems that have wide-ranging impacts across multiple domains, including defence.

When constructing a scenario, relevant documents will need to be reviewed to determine the most appropriate framework for ensuring responsible AI development and use, particularly when dealing with high-risk applications such as those in defence. Section 4 contains suitable sources from the civil sector and can serve as a source for this step.

3.2.2. Technical Paradigms

In selecting the technical paradigm for AI, the approach taken depends heavily on the scenario's requirements and the specific technological needs. DevOps and MLOps are two paradigms often considered in the context of AI development and deployment:

- DevOps focuses on collaboration between software development and IT operations, streamlining software delivery, improving testing, and reducing lead time for new features. While this paradigm might suit general software systems, its application in AI is more limited.
- MLOps (Machine Learning Operations), on the other hand, extends DevOps principles to machine learning systems. This paradigm emphasizes automation in data management, model training, deployment, monitoring, and updates. In defence-related scenarios, where models may require frequent updating to cope with dynamic data inputs or security concerns, MLOps provides more relevant capabilities.

Choosing the correct paradigm will be vital in ensuring continuous integration and delivery (CI/CD) pipelines, scalable infrastructure, and a feedback loop for model performance and risk management.

3.2.3. Viewed through the lens of norms and values of the EU (Art. 2 Treaty on EU).

In the context of AI in defence, it is critical that AI systems not only meet technical requirements but also align with these fundamental values. For example:

- Human Dignity and Autonomy: AI systems should augment human decision-making rather than replace it, particularly in sensitive scenarios like military applications.

- Equality and Non-discrimination: AI systems must ensure fairness and avoid biased outcomes that might disproportionately affect certain groups, especially in life-critical defence scenarios.
- Rule of Law and Accountability: Defense-related AI systems should be unequivocally linked to a designated responsible party—whether an individual, organization, or governmental entity—who bears full legal and ethical accountability for the system's actions, decisions, and consequences. This accountability framework ensures clear lines of responsibility, facilitates proper oversight, and helps maintain human control over critical AI-driven defence operations, with clear legal frameworks ensuring that their use adheres to national and international laws.

Embedding these values into the development and operationalization of AI ensures that technological advancements do not compromise foundational societal principles, particularly in defence contexts where ethical considerations are paramount. The findings of Chapter 0 directly underpin this point.

4. Standards and regulations for AI in Public Sector

In recent years, the rapid advancement of AI technologies has spurred widespread discussions about the need for regulatory frameworks to govern their development and deployment. These discussions have become increasingly urgent as AI applications become more prevalent and integrated into various aspects of society, raising concerns about their lawfulness, ethical implications, and robustness.

At the forefront of these efforts are regulatory bodies at both the European Union (EU) and international levels. These organizations face unique challenges and priorities in shaping the regulatory landscape for AI. For instance, they must balance the need for innovation and economic growth with the need to protect citizens from potential harm.

Meanwhile, industry stakeholders have also expressed interest in defining best practices, guidelines, and standards for the development of AI-based products. International standardization groups, comprised of representatives from industry, academia, and legislative authorities, are working together to propose standards that are compliant with regulations. The goal is to ensure that AI-based applications can be developed and implemented in a **responsible** and **trustworthy** manner, while meeting the necessary legal requirements.

The following subsections present the relevant activities addressing the definition of AI regulations, standardization, and trustworthiness, as well as an overview of the current regulations and standards being developed by these initiatives.

4.1. Relevant Activities

Standardization and regulation play critical roles in ensuring the responsible development, deployment, and governance of AI technologies across different sectors. In Europe and beyond, various organizations, initiatives, and technical committees are working collaboratively to establish frameworks and guidelines that align AI practices with ethical, safety, and regulatory standards. These efforts, ranging from European standardization forums to global collaborations like the OECD and G7, focus on fostering innovation while safeguarding societal values and ensuring AI technologies are trustworthy. This section outlines the most significant initiatives currently shaping the landscape.

4.1.1. High Level Forum on European Standardization

The High-Level Forum on European Standardization is an initiative by the European Commission aimed at enhancing the development and adoption of European standards. It brings together representatives from various sectors, including industry, academia, standardization bodies, and EU member states, to discuss and coordinate strategies for standardization in key areas such as AI, digitalization, and green technologies. The forum aims to ensure that European standards remain competitive globally, support innovation, and reflect European values and regulatory frameworks.

4.1.2. CEN-CENELEC JTC 21 “Artificial Intelligence”

CEN and CENELEC have established the Joint Technical Committee 21 (JTC 21) "Artificial Intelligence" to develop and adopt standards for AI and related data. JTC 21 will also provide guidance to other technical committees dealing with AI-related topics. The committee is currently working on European standards that will provide manufacturers with a presumption of conformity to the EU Artificial Intelligence Act [1].

4.1.3. Hiroshima Process

The Hiroshima Process refers to the outcomes and initiatives stemming from the G7 Hiroshima Summit, which took place in May 2023 in Hiroshima, Japan. During this summit, G7 leaders addressed various global challenges, including the governance and regulation of artificial intelligence. The Hiroshima Process involves the collaborative efforts of G7 nations to develop and implement frameworks and policies that ensure the responsible development and deployment of AI technologies. This includes fostering international cooperation on AI standards, ethical guidelines, and regulatory approaches to mitigate risks and enhance the benefits of AI.

4.1.4. OECD

The Organisation for Economic Co-operation and Development (OECD) is an international organization comprising 38 member countries, focused on promoting policies that improve economic and social well-being globally. In the context of AI, the OECD has been a leading entity in developing principles and guidelines for trustworthy AI. The OECD AI Principles, adopted in 2019, are designed to promote the ethical, safe, and human-centric development and use of AI technologies. These principles cover areas such as transparency, accountability, and fairness, and they provide a framework for governments and organizations to align their AI practices with societal values.

4.1.5. NIST Trustworthy and Responsible AI

The National Institute of Standards and Technology (NIST), part of the U.S. Department of Commerce, plays a crucial role in the standardization of emerging technologies, including artificial intelligence (AI), in the United States. NIST has developed a framework for Trustworthy and Responsible AI, focusing on key aspects such as:

- Risk Management: Identifying and mitigating risks associated with AI systems.
- Trustworthiness: Ensuring AI systems are reliable, secure, and operate as intended.
- Explainability: Making AI decisions transparent and understandable to users.
- Adversarial Vulnerabilities: Addressing threats from adversarial attacks on AI systems.
- NIST's guidelines and publications help organizations implement AI systems that are safe, reliable, and aligned with ethical standards, such as NIST AI RMF 100-1 [25] and NIST AI RMF 600-1 [27].

4.1.6. EASA

The European Union Aviation Safety Agency (EASA) developed EASA AI Roadmap 2.0 [28] that outlines the agency's vision for the integration of artificial intelligence (AI) in aviation. The roadmap identifies several areas where AI could have a significant impact, including aircraft design, manufacturing, maintenance, and operations. It also outlines the challenges and risks associated with the use of AI in aviation, and it sets out a framework for addressing these challenges and risks.

The EASA AI Concept Paper Issue 2 [29], on the other hand, builds on the AI Roadmap and provides more detailed objectives for the integration of AI in aviation. The objectives are organized into four categories: AI trustworthiness analysis, AI assurance, Human factors for AI, and AI safety risk mitigation. The Concept Paper is intended to stimulate discussion and debate on the role of AI in aviation, and to provide a foundation for the development of a regulatory framework for AI in aviation.

4.1.7. EUROCAE WG-114/ SAE G34

EUROCAE WG-114 and SAE G-34 represent a joint standardization initiative aimed at creating standards and guidance materials for the development and certification/approval of AI-based airborne and ground systems. This group consists of industry experts, regulatory bodies, and academic researchers, operating with the sponsorship of EUROCAE (European Organization for Civil Aviation Equipment) and SAE International (Society of Automotive Engineers).

The group has already published the "Artificial Intelligence in Aeronautical Systems: Statement of Concerns" (ER-022 / AIR6988 [30]) and is finalising the "Recommended Practice for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing Machine Learning" (ED-324 / ARP6983 [31]). The ED-324 / ARP6983 will provide comprehensive guidance on the interfaces between system, machine learning constituent, and software/hardware development lifecycles. It will also address data management, machine learning model design, validation, verification, and implementation, as well as dedicated certification activities for the Machine Learning Constituent. The first issue of ED-324 / ARP6983 will specifically focus on offline supervised learning.

4.1.8. FCAS The Responsible Use of Artificial Intelligence in FCAS Whitepaper

The **Future Combat Air System (FCAS)** is a collaborative defence project involving France, Germany, and Spain, aimed at developing a next-generation air combat system. The **Responsible Use of Artificial Intelligence in FCAS Whitepaper** [32] outlines the ethical, technical, and operational principles for integrating AI into FCAS. It emphasizes the responsible use of AI to enhance combat capabilities while adhering to ethical guidelines, international laws, and safety standards. The whitepaper addresses issues like human oversight, transparency, and accountability in the use of AI within military systems.

4.1.9. EICACS EDF Project

The **European Initiative for Collaborative Air Combat (EICACS** www.eicacs.eu**)** [33] is a project funded by the **European Defence Fund (EDF)**. It aims at ensuring the interoperability of future air combat systems, manned or unmanned platforms, legacy platforms, and their evolution (including sensors and effectors). It will promote or develop design rules and interoperability standards and assess questions regarding the implementation of AI technologies (as enablers) for military qualification, certification, or sovereignty purposes.

4.1.10. AI4DEF (AI for Defence)

The AI4DEF project [34], funded by the European Defence Industrial Development Programme (EDIDP), aims to demonstrate the benefits of AI for military applications in areas like situation awareness, decision-making, and planning optimization. To develop AI trustworthiness by design, the project combines expertise in AI and defence security, through a defence specialization of the Assessment List for trustworthy AI (ALTAI [35]). The project also utilizes a European AI LabStore concept, bringing together diverse skills and capabilities to strengthen European sovereignty in AI technology for defence systems.

4.2. AI Regulations

While the EU has emerged as a trailblazer in AI regulation with its ambitious and comprehensive approach, the international community is also working towards establishing regulations to guide the responsible use of AI on a global scale. By examining the regulatory initiatives and priorities at both levels, we can gain valuable insights into the evolving landscape of AI governance and its implications for technology development and societal impact.

An overview of existing regulatory efforts within the EU and on an international level is presented in Figure 2**Error! Reference source not found.**.

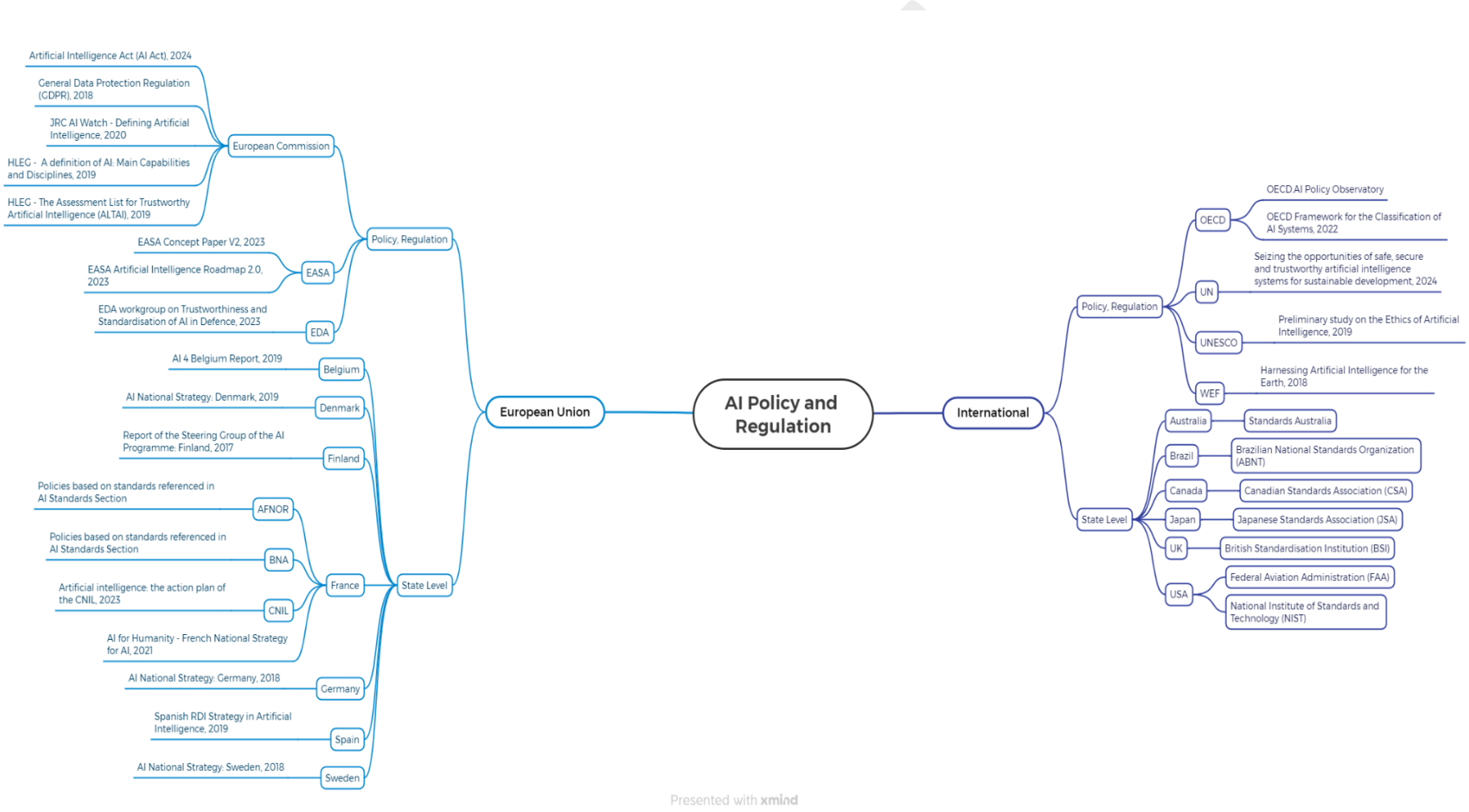


Figure 2 - AI Policy and Regulation

An example of regulation is the EU AI Act, which entered into force on August 1, 2024, establishing a regulatory framework for AI applications within the European Union. The EU AI Act adopts a nuanced risk-based approach, which ensures that regulatory requirements are proportional to the identified risks, thus maintaining a balance between innovation and safety. The EU AI Act covers a wide range of AI applications, but it is not mandatory for military applications and research, for example.

Navigating the intricate landscape of AI regulation becomes particularly challenging when considering its application in the defence sector. Unlike many civilian applications where the focus is primarily on consumer protection, privacy, and ethical concerns, including GDPR compliance, the defence sector operates within a distinct set of parameters shaped by national security imperatives, geopolitical considerations, and the complexities of modern warfare.

One of the key challenges lies in reconciling the need for stringent regulation with the imperative of maintaining military competitiveness and operational effectiveness. Regulations intended to ensure transparency, accountability, and fairness in AI decision-making processes must be balanced against the requirement for secrecy, autonomy, and rapid decision-making in military contexts. Striking this balance requires careful consideration of how to implement regulatory frameworks that safeguard against the risks of AI misuse or abuse while preserving the agility and strategic advantage of defence operations.

Additionally, the global nature of defence activities introduces complexities in harmonizing regulations across different jurisdictions. Military alliances and coalitions involve multiple stakeholders with divergent regulatory frameworks, legal traditions, and national security priorities, making it challenging to establish uniform standards for AI governance in the defence sector. Moreover, the dynamic nature of security threats and technological advancements necessitates adaptable regulatory frameworks capable of addressing emerging risks and vulnerabilities in real-time.

Furthermore, the sensitive nature of defence-related AI technologies, including autonomous weapons systems and cyber warfare capabilities, raises ethical concerns regarding their potential humanitarian impact and compliance with international law. Crafting regulations that effectively address these concerns while preserving military capabilities requires a nuanced understanding of the complex interplay between technology, ethics, and security and/ or safety.

Overall, adapting AI regulations to the defence sector entails navigating a myriad of technical, legal, ethical, and geopolitical challenges. Success in this endeavour requires close collaboration between governments, regulatory bodies, military organizations, industry stakeholders, and civil society to develop robust and agile regulatory frameworks that promote the responsible use of AI while safeguarding national security interests and upholding international norms and values.

Effective governance is paramount in the era of AI regulation. EU AI Act [1], demands rigorous Data and AI Governance practices as critical components in complying with AI regulations. Defence organizations must implement robust frameworks to ensure responsible data management, algorithmic transparency, and ethical AI development. Key aspects include implementing robust data quality measures, establishing clear roles and responsibilities throughout the Data and AI lifecycle, conducting thorough risk assessments, and aligning Data and AI practices with existing data protection laws like GDPR. By prioritizing governance, defence entities not only comply with regulations but also build trust in their AI systems, fostering innovation while mitigating risks.

For further detail, a table with full list of regulations identified so far can be found at “TAID Annex 06 Standards & Regulations”.

4.3. AI Standards

The standardization of trustworthiness in AI is an evolving field that involves the development of consensus-based guidelines, methodologies, and best practices to promote responsible AI development and deployment. Central to this effort is the implementation of robust Data and AI Governance frameworks, which provide the foundation for ensuring ethical, transparent, and accountable use of data and AI systems throughout their lifecycle. These governance structures encompass policies, processes, and controls that guide the collection, storage, processing and use of data, as well as the design, development, and deployment of AI models.

By establishing common standards and benchmarks through comprehensive Data and AI Governance practices, stakeholders can mitigate risks, enhance interoperability, and foster innovation while upholding ethical principles and societal values. In this context, exploring the trustworthy AI regulatory and standardization frameworks, underpinned by strong governance mechanisms, offer valuable insights into the challenges, opportunities, and implications for advancing the responsible use of AI in an increasingly interconnected and technology-driven world. Effective Data and AI Governance not only ensures compliance with evolving standards but also builds trust among stakeholders, thereby facilitating the widespread adoption and acceptance of AI technologies.

Figure 3 shows a representation to categorize the current trustworthy AI standards (green flag for published standards, orange one for standards under progress and red flag for stand-by or just initiated standards).

The CEN-CENELEC Joint Technical Committee (JTC 21) proposed a roadmap for AI standardisation, which was evaluated by the European Commission's Joint Research Centre. The evaluation identified many gaps in existing international standards and suggested additional standards to support the AI Act [1].

JTC 21 has already adopted some AI Harmonised Standards, and CEN and CENELEC have published a work programme detailing progress on developing additional standards. However, the completion of Harmonised Standards is expected to be delayed until late 2025, potentially leaving companies with less time to implement them before the AI Act's enforcement in August 2026.

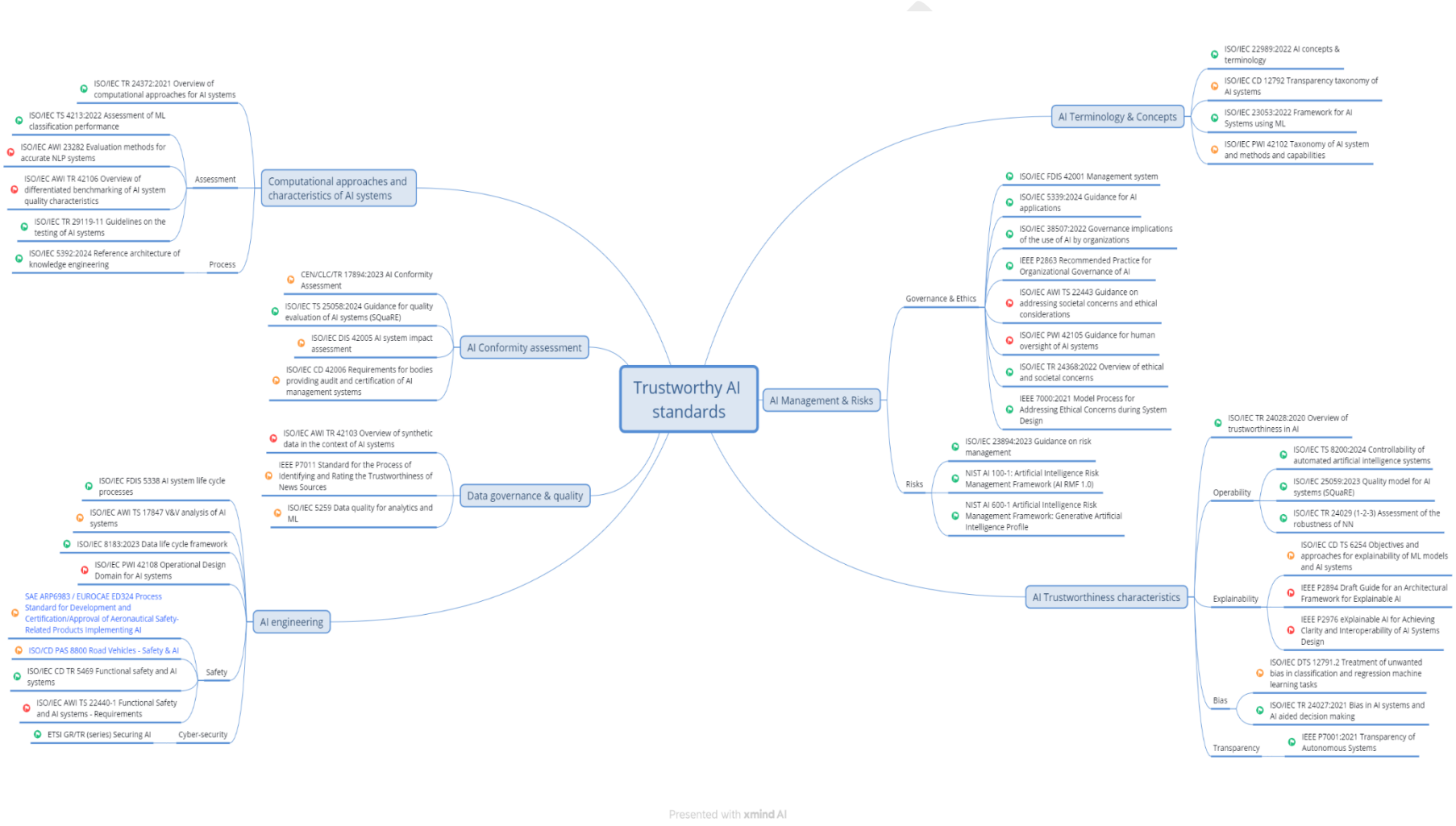


Figure 3 - Mind map of standards for trustworthy AI

This figure emphasizes that:

- The majority of the listed standards are horizontal. Vertical ones (like the ones mentioned in blue) need to be developed as references for given sectors.
- A lot of standards are still in progress for most of categories (except governance) although some of them are published every year.
- Some given areas like terminology, guidance or trustworthiness characteristics seem to be well covered but have still to be refined for military purpose to address topics like autonomy or risks (considering missions, not only safety[3]).
- Missing Defence-specific constraints have also to be explored like data frugality or incremental qualification.

Nevertheless, it does not show that most of trustworthy AI standardization initiatives are still restricted to supervised ML till now. Defence needs to consider other kinds of learning like Reinforcement Learning and other types of AI like hybrid or cognitive AI to find the best trade-off for efficiency purpose.

For further detail, a table with full list of standards identified so far can be found at “TAID Annex 06 Standards & Regulations”.

[3] A positioning on safety and mission critical systems is available in the Appendixes.

5. Testing and Evaluation: AI Trustworthy Engineering Lifecycle

Evaluating the trustworthiness of AI systems should result from a standardized process, independent of the technologies involved and the specific application. Nevertheless, it appears mandatory to consider the peculiar operational environment applied to defence applications. In the effort to define a guideline for European MoDs, the section 5.1 proposes a draft process for the acquisition of AI-empowered systems for defence. The section 5.2 describes a thorough AI trustworthy engineering lifecycle that is applicable to different families of AI technologies (machine learning, symbolic AI, hybrid AI). Then, the importance to manage frequent upgrades of AI-based systems is discussed in the section 5.3 dedicated to incremental development and qualification. In addition, section 5.4 presents the introduction of toolkits that can support the implementation of this AI trustworthy engineering lifecycle with a special focus on verification and validation technics. Finally, the chapter ends with section 5.5 which provides an overview of Testing and Evaluation in Defence function and how it can be applied to AI-based systems.

5.1. Acquisition Process

Following the NATO and national MoDs approach to acquisition processes, it is expected to include specific requirements expressed in terms of technical properties in the technical specification for AI-empowered systems. Considering the potential risks related to the use of AI in the military environment, a technical risk assessment should be conducted before the *critical design review*, allowing the military personnel to evaluate the potential impacts on the tactical and operational capability of the system.

The expected effects of risk evaluation completed before the critical design review (see Figure 4) is precisely part of the design of the system. For each risk evaluated, mitigation strategies should be implemented in the overall system design.

The effectiveness of the mitigations is then evaluated during the verification and validation phase as part of the system acceptance or the system qualification process. To measure the effectiveness of each mitigation applied, relevant metrics should be indicated for any risk evaluated to better define appropriate trial procedures and help stakeholders in the residual risk evaluation process.

As general guidelines we provide three annexes to the present white paper, which provides:

- Risk Analysis: relevant examples of risks to be evaluated and related properties.
- Trustworthiness Properties: properties to use as metrics in the test and evaluation process.
- Toolkits and frameworks: examples of toolkits for risk assessment.

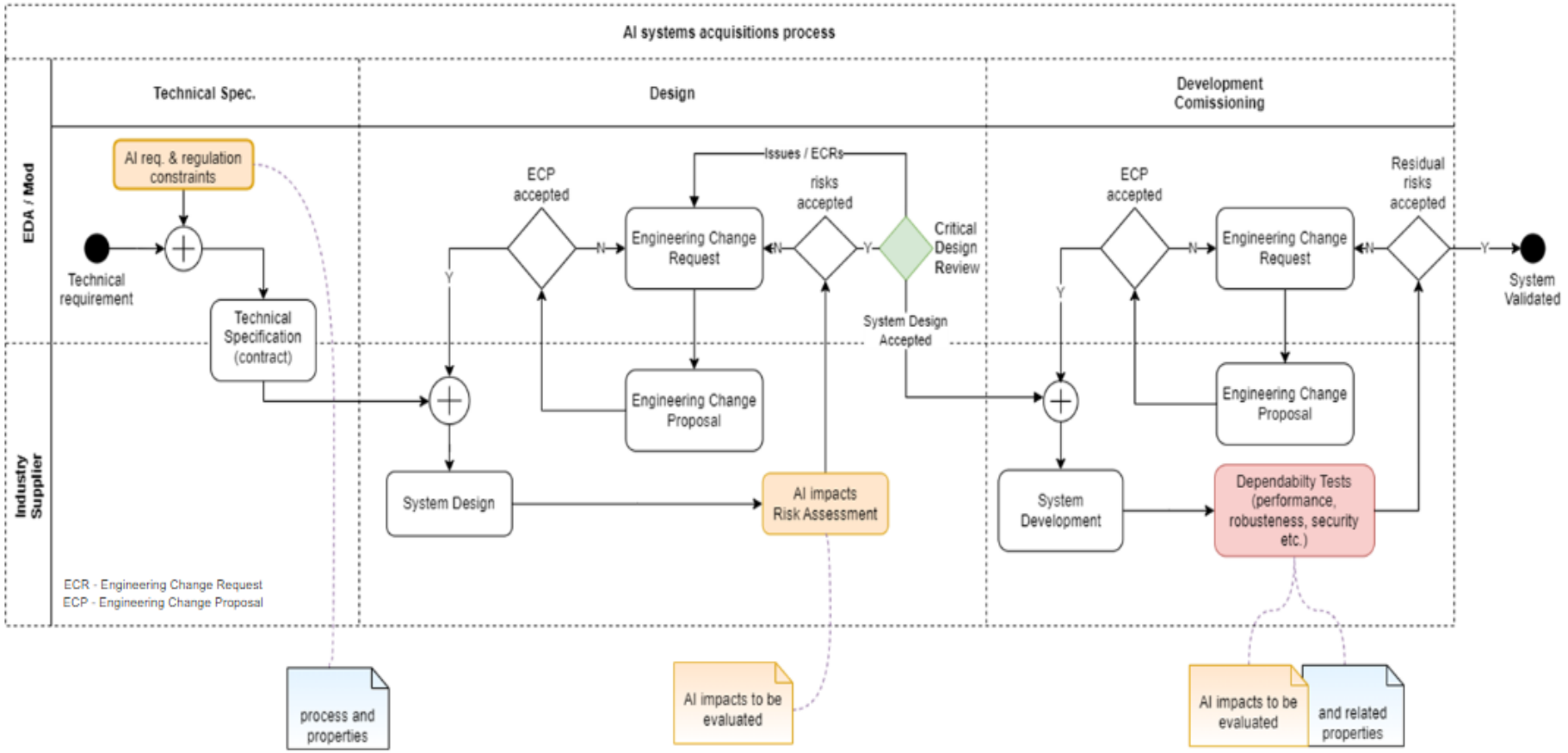


Figure 4 - Proposed acquisition process

5.1.1. Risk assessment

The risk assessment phase is crucial for the design of a trustworthy AI system. In order to help build the risk assessment document, the dedicated "TAID Annex 01 Risk Analysis" provide a basic set of risks from which one can start to build its own specific one.

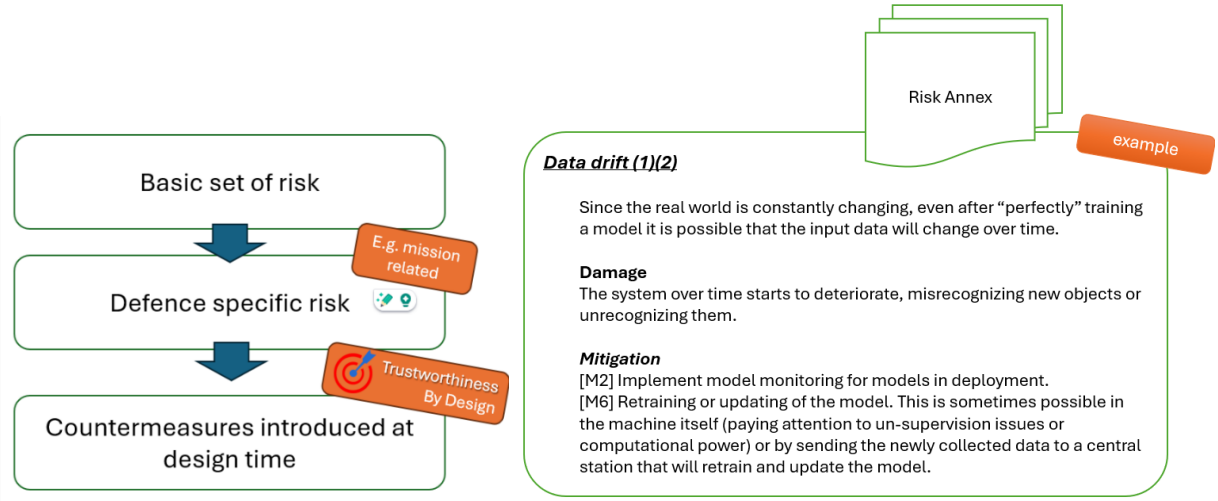


Figure 5 - Risk assessment objective

Stakeholders should be able to understand the impacts and eventually accept the residual risk. To do that, the selection and monitoring of trustworthiness properties is crucial. The way in which we can measure them and the direct relation to risks could be used to evaluate the quality of the design before the critical design review.

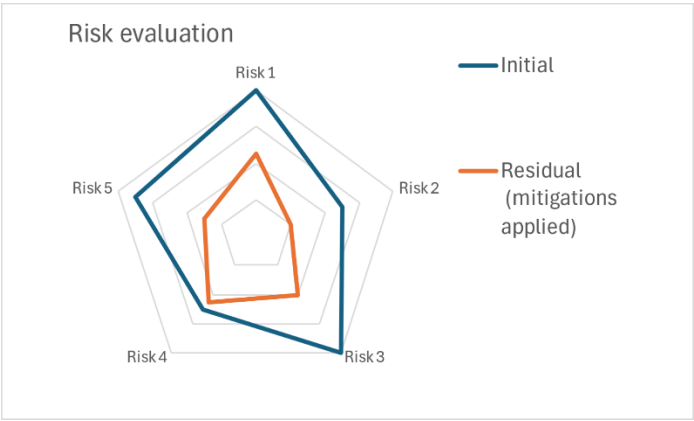


Figure 6 - Example of residual risk evaluation

Thus, each risk is related to one or more properties whose metrics could be used to measure the effectiveness of the mitigation strategies applied at design time.

CONTROLLABILITY [PR-02]

References: EN ISO/IEC 22989

Definition: property of an AI system that allows a human or another external agent to intervene in the system's functioning. The ability to control and manipulate inputs, conditions, or parameters during testing so that you observe specific behaviors or evaluate specific responses within specific contexts.

Metrics: check ISO/IEC TR 24028:2020 and ISO/IEC TR 8200:2024

Figure 7 - Property detail example

The same properties will be used to also evaluate the implementation of AI-based systems during the trials and throughout the system lifecycle. A list of properties and their definitions is presented in the following Table 1. For a more detailed approach (including risks correlation and metrics) see "TAID Annex 02 Trustworthiness Properties".

Table 1 - List of properties for evaluation of risks due to integration of AI technology in defence systems

Property ID	Property Name	Property Description
TAID-01	Accountability	State of being answerable for actions, decisions and performance.
TAID-02	Accuracy	The degree to which models and data have attributes that correctly reflect the true value of the intended attributes of a concept or event in a particular context of use.
TAID-03	Resilience	Resilience is the ability of the system to recover operational condition quickly following an incident.
TAID-04	AI self-protection	Integrated features and increased capacity of the AI to prevent non-intended third-party interactions like disclosure, reverse engineering, and miss-usage.
TAID-05	HW capacity to support AI complexity	The ability of the HW to support the execution of an AI algorithm/application in a safe and efficient way.
TAID-06	Autonomy	Autonomy is the ability of a system to achieve goals while operating independently of external control. For defence, it means facing potential intentional and unintentional challenges that put the mission at risk.
TAID-07	Availability	Being accessible and usable on demand by an authorized entity.
TAID-08	Data Completeness	Degree to which a data set sufficiently (according to specified criteria) covers the operational design domain for the intended application.
TAID-09	Confidentiality	Information is not made available or disclosed to unauthorized individuals, entities, or processes.
TAID-10	Consistency	Degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities.
TAID-11	Controllability (incl. Meaningfull Human Control)	Property of an AI system that allows a human or another external agent to intervene in the system's functioning. The ability to control and manipulate inputs, conditions, or parameters during testing so that you observe specific behaviors or evaluate specific responses within specific contexts.
TAID-12	Explainability	Aspects including data provenance and the ability to provide an explanation of how an AI system's output is determined. It is important to have it clear why an AI algorithm took a certain decision not only to understand the important factors that led to that decision but also to generate more trustworthiness in the system itself from the user perspective.
TAID-13	Function gain, extension	The usage of the AI component/technology produces a function gain, or an extension of existing function(s) originally developed without AI.
TAID-14	Generalisation	Generalization is the ability of ML models to provide accurate outputs when fed with inputs not seen during the training phase.

Property ID	Property Name	Property Description
TAID-15	Governability	AI applications will be developed and used according to their intended functions and will allow for: appropriate human-machine interaction; the ability to detect and avoid unintended consequences; and the ability to take steps, such as disengagement or deactivation of systems, when such systems demonstrate unintended behaviour.
TAID-16	Homologation/certification	The processes followed to homologate or certify a system as preconditions to release it for operation.
TAID-17	Data Integrity	It refers to the assurance that data and its values remain unaltered and uncorrupted throughout the processes of collection, storage, and processing.
TAID-18	Interpretability	Capacity for an external observer to understand the internal behaviour of an AI system and find its meaning.
TAID-19	Maintainability	Measure of how easy it is to keep a software system running smoothly and effectively. A maintainable system can be easily adapted to changing needs, whether those changes are made by the original developers or by new members of the team.
TAID-20	Predictability	Property of an AI system that enables reliable assumptions by stakeholders about the output.
TAID-21	Recognition	Automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories. Source: Pattern_Recognition_and_Machine_Learning
TAID-22	Reliability	Property of consistent intended behaviour and results that enables to provide required prediction, recommendation, and decision consistently correctly during its operation stage.
TAID-23	Repeatability	Measurement precision under the condition of replicate measurements within a short period of time, with the replicate measurements made using the same operator, location, and measuring equipment.
TAID-24	Representability	A data set is representative when the distribution of its key characteristics is similar to the actual input state space for the intended application.
TAID-25	Reproducibility	Degree to which an AI model can be reproduced using the same inputs (data or knowledge or both) and the same engineering processes, activities and tools, thereby obtaining exactly the same or similar results according to specified similarity criteria.
TAID-26	Responsibility	Obligation to act and take decisions to achieve required outcomes.
TAID-27	Reusability	Increased possibilities for reusage of the AI technology (or the system that integrates it) under larger/new operational conditions.
TAID-28	Robustness	Ability to maintain their level of performance, as intended by their developers, under any circumstances. (Local Robustness - Stability)
TAID-29	Sovereignty	The deployment of technology encourages or ensures proper sovereignty for state members, and/or EU.
TAID-30	Specifiability	Extent to which the AI constituent can be correctly and completely described through a list of requirements.
TAID-31	Stability	Stability of the learning algorithm refers to ensuring that the produced model does not change a lot under perturbations of the training data set.
TAID-32	Sustainability	More efficient usage of the energy budget of the system itself or in the energy spent for system design, production, supply chain, etc.
TAID-33	Testability	Degree of effectiveness and efficiency with which test criteria can be established for a model based on its ODD and tests can be performed to determine whether those criteria have been met.
TAID-34	Timeliness	Extent to which data from a source arrive quickly enough to be relevant.
TAID-35	Traceability	Capability to keep track of the system data, events, during the development, deployment, operation processes, and decommission.
TAID-36	Transparency	Communicating appropriate information about the system to stakeholders (e.g. goals, known limitations, definitions, design choices, assumptions, features, models, algorithms, training methods and quality assurance processes). Additionally, transparency of an AI system can involve informing stakeholders about the details of data used (e.g. what, where, when, why data is

Property ID	Property Name	Property Description
		collected and how it is used) to produce the system and the protection of personal data along with the purpose of the system and how it was built and deployed. Transparency can also include informing stakeholders about the processing and level of automation used to make related decisions.
TAID-37	Usability	Increased possibilities for the usage of the AI technology (or the system that integrates it) under predefined operational conditions.
TAID-38	Observability	Observability is a measure of how well internal states of a system can be inferred from knowledge of its external outputs.
TAID-39	Quality	The OECD Quality Framework is built around eight considerations: 1. Relevance 2. Accuracy 3. Credibility 4. Timeliness 5. Accessibility 6. Interpretability 7. Coherence 8. Cost-efficiency
TAID-40	Causality	Ability to establish causal relationship between events to ensure the fair behaviour of systems.
TAID-41	Model Correctness	Ability of a model to maintain its level of performance under all nominal (not processed by the model robustness) conditions within the ML Operational Design Domain.
TAID-42	Dependability	Ability to perform as and when required.
TAID-43	Recoverability	Capability of a product in the event of an interruption or a failure to recover the data directly affected and re-establish the desired state of the system.
TAID-44	Bias	Systematic difference in treatment of certain objects, people or groups in comparison to others.
TAID-45	Data Balance	In a balanced dataset, each class contributes equally to the overall composition.

Traditionally, risk-based approaches can be used for safety-critical applications with the definition of assurance levels according to the severity of the consequences of a failure. The EU AI Act (which explicitly excludes military use cases) is also based on a risk-based approach, which strengthens the relevance of risk-based approaches. Other ISO standards like the ISO/IEC/IEEE 29119 series or currently under development are also based on a risk-based approach.

5.2. End-to-end life cycle for AI

5.2.1. AI Engineering Lifecycle: Machine Learning and Symbolic AI System Engineering

The engineering lifecycles of machine learning (ML) and symbolic AI (also known as knowledge-based AI) share certain commonalities but also diverge on specific aspects due to the intrinsic differences in their approaches to AI. Both lifecycles rely on engineering processes that combine AI-specific methodologies with traditional engineering practices at system level, as well as at software and hardware implementation level. Understanding these lifecycles provides valuable insights into how AI systems are developed, from design to deployment.

5.2.2. System Engineering

Both ML and Symbolic AI engineering lifecycles begin with system engineering that should adhere to existing system standards and practices, since using AI can be seen as an implementation choice driven by the fact that data is available, or knowledge exists. This foundational stage involves system requirements elicitation, the definition of system

architecture, and specific risk analyses such as safety, security, and human performance assessments. Examples of system standards commonly used in Europe to develop military systems are NATO Standards (STANAGs, AQAPs), UK Defence Standards (DEF STAN), United States Military Standards and Specifications (MIL-STD and MIL-SPEC), ISO/IEC/EUROCAE/IEEE standards and guidelines, etc.

5.2.3. AI W-shaped Development Lifecycle

The development of both ML and Symbolic AI systems can be visualized as a W-shaped lifecycle [EASA AI Concept Paper issue 2 [29] and EUROCAE ED-324/SAE ARP-6983 [31]]. This model expands on the traditional V-model by adding layers that address the specificities of AI constituent development. The term “AI constituent” (inspired from EUROCAE ED-324/SAE ARP-6983) is defined as the combination of an AI model (ML or Symbolic) with its necessary pre/post (data or knowledge bases) processing. There are strong links between the AI model and its pre/post processing making them not separable for verification (test) purposes (i.e., it is not possible to verify the AI model without its pre/post processing).

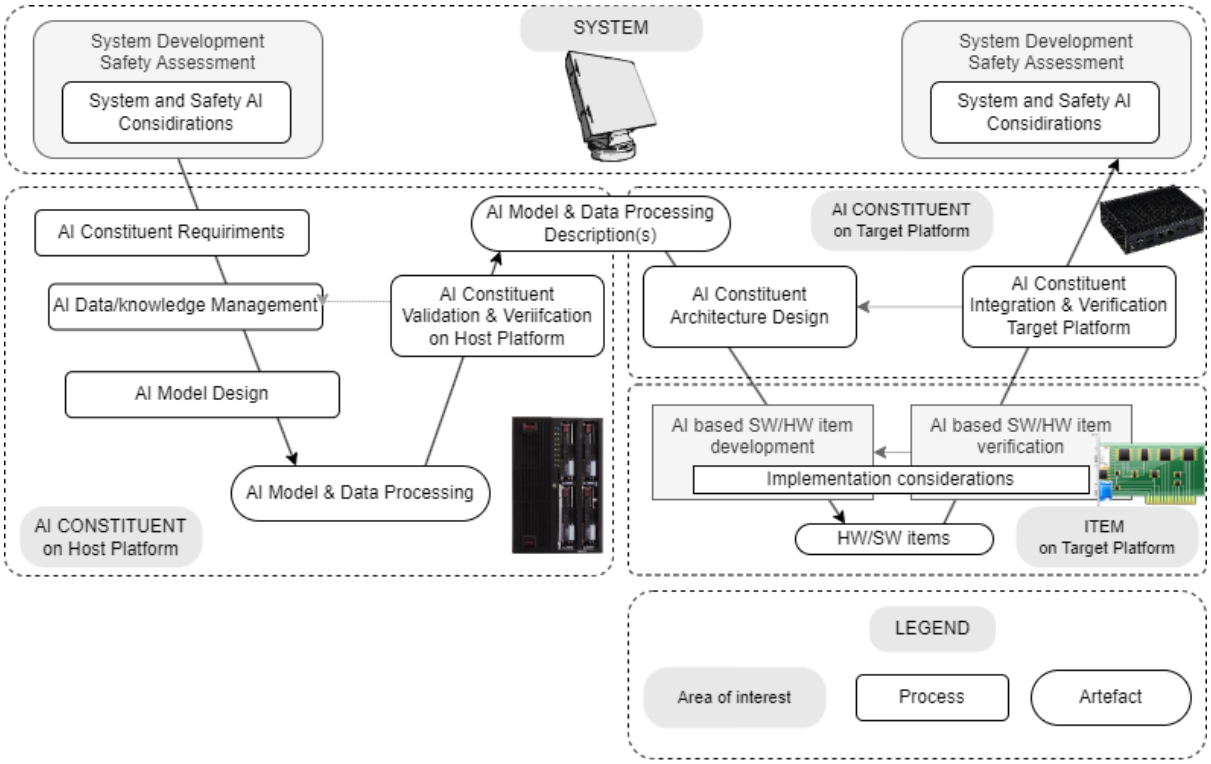


Figure 8 - AI W-Shape lifecycle for ML and Symbolic AI

The W-Shape can be divided in two parts: the engineering processes performed on the Host Platform (e.g. on a private computing cluster, on a secured partition on the cloud), and the engineering processes performed on the Target Platform (e.g., specific HW embedded in a ground or aerial vehicle).

5.2.4. W-Shape Engineering Processes on the Host Platform

AI Constituent Requirements

The lifecycle commences with a detailed refinement of the system requirements (including the description of the operating environment/operational domain) into more detailed AI constituent requirements. This phase includes the characterization of the Operational Design Domain

(ODD) of the AI constituent, the description of its logical architecture, and if needed the description of scenarios.

AI Data/Knowledge Management
A critical juncture for both ML and Symbolic AI systems, this stage involves the systematic organization, storage, and retrieval of data or knowledge. For ML, it entails managing datasets (e.g. planning for data source identification, collection, data preprocessing and feature extraction) required for training, validation and test (independent verification). Conversely, for Symbolic AI, it focuses on the structured representation of domain knowledge and the mechanisms for its manipulation and access.

AI Model Design
For ML, this design stage involves the selection and settings of learning algorithms, defining ML model architectures, iterations of training and evaluation the candidate ML model, and if needed optimization of the size of the trained ML model (pruning, quantization, etc.). For Symbolic AI, it entails the formulation of knowledge bases, inference rules, and logic structures.

AI Constituent Validation & Verification on Host Platform
Validation and verification (V&V) processes are pivotal, ensuring that the AI components meet their allocated requirements, including relevant trustworthiness properties such as stability, robustness, generalisation, reproducibility, etc. This stage employs a variety of techniques, from simulation and testing against curated datasets for ML constituents to formal methods for verifying the logic and consistency of Symbolic AI constituents.

5.2.5. W-Shape Engineering Processes on the Target Platform

AI Constituent Physical Architecture Design
This step enables to move from the logical architecture of the AI constituent to its physical architecture on the target platform. It involves decisions on the decomposition of the logical architecture into many software and hardware items, choices of hardware and network configurations, and the integration interface with existing software/hardware items, subsystems, or systems.

AI-based SW/HW Item Development
This step can use existing software and hardware development guidelines (such as ED-12C/DO-178C [36] and ED-80/DO-254 [37]). For ML, this may include the implementation of optimized ML models in software, tailored for specific hardware accelerators. In Symbolic AI, it involves coding the knowledge representation and inference engines in software, alongside any specialized hardware for logic processing.

AI-based SW/HW Item Verification on the Target Platform
Prior to full integration, AI-based SW/HW items undergo targeted verification on the target platform. This ensures compatibility, performance, and reliability within the intended ODD, employing automated testing frameworks, analysis, and reviews.

AI Constituent Integration and Verification on Target Platform
The culmination of the lifecycle is the integration of the software and hardware items of the AI constituent on the target platform, followed by comprehensive system-level V&V. This phase ensures that the AI constituent, once integrated, operates as intended and meets its allocated performance requirements on the target platform.

5.2.6. Hybrid AI Engineering Lifecycle

The distinct engineering lifecycles of Machine Learning (ML) and Symbolic AI, as delineated above, can be combined into a hybrid AI engineering lifecycle that leverages the strengths of both paradigms to create more robust and efficient AI systems. This hybrid approach integrates the data-driven flexibility and learning capabilities of ML with the explicit reasoning and domain-specific knowledge representation of Symbolic AI.

By doing so, it aims to amplify the advantages - such as the ability of ML to manage complex, high-dimensional and unstructured data and the explainability/interpretability of Symbolic AI - while mitigating their respective drawbacks and risks, such as the opaqueness of ML models and the rigidity of Symbolic AI constituents. Such hybrid approach enables the development of AI-based solutions that are not only more powerful and effective but also more trusted by end users (operators, certification authority, etc.), addressing specific challenges raised by the introduction of AI in the Defence domain.

5.3. Incremental development and qualification

This section discusses the need for defence use cases to progressively update AI models (e.g. to handle enemy manoeuvres) at a much faster rate than for traditional civilian use cases (e.g. a model updated overnight to prepare for the next mission the next day) throughout appropriate change management and risk management procedures. This is an important specificity of the defence AI field that is enabled by the high level of automation of AI development technologies.

This challenge has an impact on the development cycle and the way in which updates are qualified, as going through all the stages of the initial cycle again could be incompatible with operational time constraints. New approaches need to be invented, involving the authorities in charge of the approval/certification of the systems, to formalize the necessary trade-off between the need to ensure the safety of the system and its users, on one hand, and the strategic interest of the mission and war aims, on the other. This challenge is illustrated in the Figure 9 for machine learning, where the pace of change is correlated with the temporal validity of the data used for training.

Figure 9 - Camouflage of anti-aircraft missile batteries (generated with AI tool)

During the night, the enemy camouflages anti-aircraft missile batteries with vegetation, that causes changes in the way the AI model has learnt to distinguish background data (e.g., vegetation in an image) from data of interest (here the anti-aircraft missile batteries). The detection performance of the AI model can be strongly impacted leading to increased false negatives (the anti-aircraft missile batteries are no longer detected by the AI-based system embedded on a fighter aircraft). This necessitates the decision to retrain before the next mission through a rapid incremental development and qualification process the AI model to detect and recognize camouflaged targets.

5.4. Toolkit

The advent of artificial intelligence, particularly machine learning, has led to significant changes in the world of industrial production and service delivery. The integration of these advanced technologies has opened new possibilities for optimizing and improving processes, enabling companies to achieve unprecedented levels of efficiency and productivity. However, it is important to remember that such models can be subject to manipulations seeking to deceive or evade the models themselves. Thus, there has been a growing interest in understanding how to ensure the proper functioning of the algorithms used at the core of machine learning models.

The presence of increasingly sophisticated attacks represents a significant threat to the effectiveness and security of crucial applications that must not be underestimated. Solutions for protecting the models and the data they handle are fundamental, especially in sectors where regulations are in place. For organizations to remain compliant with increasingly elaborate regulatory requirements, they must be prepared to define and apply the correct combination of approaches and methods to make their models secure.

There are several specialized products and frameworks that can be used to evaluate the robustness of machine learning models and perform validation tests that will be introduced in "TAID Annex 03 – Toolkits and frameworks". Notice that a brief discussion on the evolution of tools and standards is provided in the Appendixes.

AI systems will play an increasingly important significant role in future military applications. As AI systems differ from existing rule-based algorithms and systems, it is important that AI systems are also trustworthy for future users. A key aspect of creating trust is the systematic validation of AI systems under relevant operating conditions what requires **product-neutral** evaluations of AI systems regarding their potential as well as their weaknesses, based on standards relevant to military systems.

5.5. Test and Evaluation of AI in defence systems

Test and Evaluation (T&E) is an engineering function which allows us to understand, even in advance, which are the operational risks, and the technical deficiencies related to the use of a new technology or to the integration of a technology into an existing defence system. This process spans the entire lifecycle of a system, from initial development and integration to decommissioning or upgrading. T&E contributes to the increase of reliability, robustness, and safety of a system, by identifying potential risks and assessing the mitigations strategies which will ensure safe, secure and efficient design, deployment and operational use of the system. For AI-based defence systems, this role becomes even more critical due to the unique challenges posed by autonomous decision-making, data-driven behaviours, and cybersecurity concerns.

There are three main T&E workstreams for AI-based systems:

1. Developing methods, procedures and standards to test and evaluate AI-based systems, including a comprehensive analysis of their benefits and risks.
2. Leveraging AI technologies to enhance the testing and evaluation process for defence systems.
3. Defining the boundaries of AI within defence systems from a Testing and Evaluation perspective to ensure safe, effective and reliable AI integration while maintaining secure system operation.

The implementation of these workstreams is critical to ensure the reliability and security of AI technologies as well as their effective application in assessing broader system capabilities.

T&E actively participates in the development of AI-based systems from the earliest stages, playing a pivotal role in the Validation and Verification (V&V) process. This involvement is integral to the qualification and, where applicable, certification of defence systems, ensuring optimal performance and compliance with the safety and regulatory requirements. The synergy between T&E and the V&V process is systematically implemented across the following six phases:

Phase 1: Concept Exploration and Requirements Definition

Objective: Establish clear and measurable objectives for AI deployment in defence applications and based on these define the appropriate requirements.

Phase 2: Data Acquisition and Model Training

Objective: Acquire data to develop robust AI models tailored to defence-specific tasks.

Phase 3: Simulation and Testing in Controlled Environments

Objective: Conduct initial T&E of AI algorithms in simulated or controlled conditions to validate the associated requirements and the AI boundaries and assess the system functionality and safety.

Phase 4: Field Testing and Live Exercises

Objective: Verify AI-based system performance and compliance with the validated requirements under real-world operational conditions.

Phase 5: Evaluation and feedback for Continuous Improvement

Objective: Refine AI models and system design based on T&E results and operational feedback to enhance reliability and effectiveness.

Phase 6: Final Validation for Deployment

Objective: Conduct comprehensive validation to ensure that the AI-based system is fully operational and in compliance with all regulatory, safety and performance requirements before deployment.

6. Human Factors

6.1. Introduction

This chapter introduces several human factors concepts in relation to Trustworthiness for AI in defence. This chapter also highlights the importance of a valued-based approach to ethics and the adoption of a systems perspective to uphold inherent values of the EU and to ensure trustworthy AI for defence, the preservation of peace and human life. It also presents a case-study to demonstrate the impact of AI on the human through the lens of a systems approach.

6.2. The relevance of Human Factors to Trustworthy AI

Human Factors is a discipline concerned with the analysis and design of sociotechnical systems to improve human wellbeing and overall system performance [38]. In other words, the application of what we know about human beings, their abilities, characteristics and limitations, to the design of equipment they use, environments in which they function and jobs they perform [39]. Human Factors encompasses knowledge from a range of scientific disciplines that supports human performance through the design and evaluation of equipment, environments and work in order to improve system performance [39].

With a view to improving the system performance of defence technologies and operations, human factors have a substantial contribution to make towards increased Trustworthiness, especially where meaningful human control, autonomy and ethical practice are concerned. AI is still in its relative infancy [40], making it crucial to avoid relying solely on humans as the ultimate failsafe for unexpected events when technology fails. Instead, fostering a collaborative intelligence between human and machine counterparts is essential to ensure robust and adaptive systems [41]. There also needs to be a thorough understanding of how all counterparts function from an integrated perspective and the nuance of the interactions between the levels of the system. It is vital to demonstrate how the humans in the system are likely to be impacted so that the wider system around them can be designed in such a way to optimise operational performance and human wellbeing.

6.3. Key requirements for Trustworthy AI

The Ethics Guidelines for Trustworthy AI [23] offers practical guidance for achieving and implementing Trustworthy AI. It identifies a set of seven key requirements that should be met, and which are based on the principles established at Chapter I of the document (Lawful, Ethical and Robust).

Note that the interrelationships of the seven requirements are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle. Additionally, these requirements apply to all stakeholders involved in the life cycle of AI systems, including those who develop technologies, those who deploy them, end-users, and the wider society.

Socio-Technical Systems Approach to Trustworthiness for AI in Defence

A Socio-Technical System approach considers the dynamic interaction between all system elements, including tasks, individuals, teams, organizations, regional/national influences, industry, and regulatory frameworks. The "socio" aspect encompasses the people within the system—their communication, collaboration, interactions, and organizational structures. The "technical" aspect refers to the technologies, tools, and infrastructure that enable the system to function effectively. This holistic perspective is essential for designing and maintaining systems that ensure safe and efficient operations.

A Socio-Technical approach addresses critical system requirements, such as training, procedures, standards, risk assessments, and incident investigations. It also identifies the skills, knowledge, and abilities needed by individuals, the tools and technologies required, and the governing rules, regulations, and laws. The interconnectedness of system levels means that changes at one level can influence outcomes across others. For instance, targeted training can enhance mission success at the task, individual, and team levels, while new standards can improve technology design and usability, enabling teams to achieve their goals more effectively.

Understanding the relationships and nuances between these levels is vital for driving meaningful improvements in safety-critical operations [42]. With the rapid advancements in technology and machine learning, modern systems have become increasingly complex, distributed, and multi-layered. These systems require a sophisticated approach to manage and synchronize "networks of many interacting and interdependent human and machine roles" [43]. This approach ensures that human-machine collaboration remains effective and adaptable to the demands of evolving technology and operational contexts.

Figure 10 depicts a socio-technical system (adapted from [44], [45]*)* and the levels contained within the system.

Figure 10 - Socio-technical Systems Levels, Adapted from MacLachlan 2017, McVeigh et al. 2022

Regulatory *(Regulations, Laws, Legal and regulatory bodies)*: Collective international agreement on how defence operations are regulated and legally held to account. International regulations are stablished in accordance with international bodies such as NATO and the UN, this includes International Humanitarian Law and other legal bodies.

Industry *(Collective Defence Force Grouping)*: Collective Standards, agreements, and practices on defence operations at the EDA/ EU level. This includes consensus on how the EDA interact with other defence forces and nations outside of EU remit.

Regional / National *(National Force)*: Organizational Concerns at the national Level i.e. defence from integrated perspective (i.e. includes air, sea, land, cyber).

Organization *(The defence force/domain)*: This represents the operational level for the particular domain (e.g. air, sea, land, cyber) within the defence forces. This includes the structures, policies, procedures, rules, regulations, provision of resources, chain of command relevant to the operational domain. Support for professional development, training, maintenance of certification, human resource management, acquisition supplies and technologies, risk management, safety management etc. The organizational culture, command structure and leadership are highly relevant to the way in which the organization functions.

Team *(The unit/ The team members)*: All the human and non-human members of the team. Human – non-human teaming considerations should be taken into consideration for all aspects of interaction between team members. This includes the ability of the team to communicate and co-ordinate together effectively to achieve mission / operational success. Team co-ordination includes decision-making, give, and follow appropriate commands and the provision of support for fellow team members. Collective competence of the team (i.e. knowledge, skills, training, experience, tacit knowledge) and team culture are inherent.

Individual *(The person)*: The human team member. This includes the person's general abilities – physical, cognitive, state of physical and mental wellbeing, interpersonal skills, etc. A clear match required between skills, demands, training, competence, and the human's ability to cope with the dynamic operational and environmental conditions and psychophysiological stress.

Task *(The Job /Mission)*: The tasks and subtasks required to complete the process. Clear understanding of the task's nature, duration, complexity, resources and demands (i.e. physical, cognitive, information, co-ordination etc.) are required. Specific skills, technologies and people are required to complete the task successfully.

The benefits of using a systems approach can be summarized as follows [46]:

- Multiple perspectives taken are into account and appropriate trade-offs are ensured.
- Functioning of the system as a whole is addressed.
- Maximizes buy-in from stakeholders and avoids placing too much emphasis on a single system level on its own.

The risks of not using systems approach can be summarized as follows [46]:

- Imbalance and ineffective or hazardous operations due to misalignment of elements in the system.
- Failure to identify potential risks and hazards ahead of time and possible emergence during system use.

Systems Approaches have been in use since the 1950s [47] and have been used increasingly across safety-critical industries since then. Human Factors research has changed perspective on how the human is perceived in the system. As can be seen in Figure 11, progression from Safety I to where we are today is welcome, however, novel technologies and systems must be designed to enable humans to be flexible, resilient and to be able to handle unexpected events.

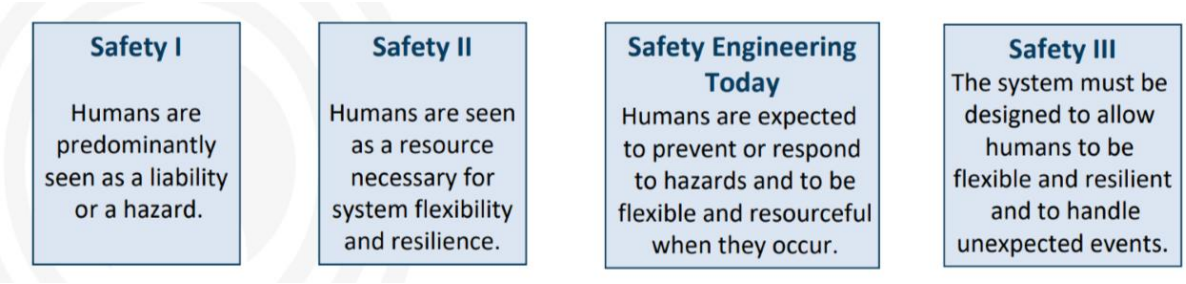


Figure 11 - Safety I to Safety III- Progressive perception of human in the system, HFES,2020

The perception of the human has progressed from being considered the sole source of error and fallibility, to viewing "human error" in the system as "socio-technical systems error" [48]. Indeed, systems are now seen as "brittle" where previously failure was attributed to human error, unpredictable parts and sub-systems [49]. AI does not function in a vacuum – it does so as a part of a large and complex socio-technical system with human agents in distributed networks throughout.

It is imperative that these safety-critical "systems of systems" are supported appropriately ([50], [51]). To fully grasp how the wider systems support the human(s), it is necessary to understand how the tools and applications function with human operators at an integrated level. The design of systems, technologies, training, evaluation, risk management frameworks etc. must also be cognisant of this.

The appendices contain a case-study to highlight and contextualize some of the challenges the Defence Community must meet to ensure Trustworthiness in AI and further recommends next steps required to do so. The system of interest is a drone with autonomous weapon capability. The case study includes the relevant stakeholders and demonstrates how AI may impact (both positive and negative) the operation, the stakeholders and the wider socio-technical system at all levels.

Open Human Factors Issues for the trustworthiness and use of AI for defence.

This section highlights some of the open human factors issues surrounding the trustworthiness and use of AI for defence.

It is imperative that a full understanding of how the technology will be used and how it will interact with the human agents (and other stakeholders) in the system is obtained. Both a detailed requirements analysis and Socio-technical Systems modelling of a system are essential if it will be possible to maximize the benefits of AI whilst keeping the humans in the system safe and able to operate at the best of their ability and capability.

Lawful – As mentioned in previous chapters, it is vital that all activities and operations adhere to both EU and international Law in relation to AI. **Ethical** – As detailed in Chapter 0, the ethical use of AI is of paramount importance if AI is to be trusted as a valued resource, support and indeed team member (i.e. non-human agent). All associated data from monitoring should also be ethical with assurance that use of said data will be in accordance with a just culture.

Robust – The associated technology and AI functionality should indeed be robust; however, both must be appropriate for the social context in which AI operates, i.e. in full support of the humans in the system, be they human agents of defence technologies, members of EDA defence forces or civilians.

AI does not function in a vacuum – it does so as a part of a large and complex socio-technical system with human agents in **distributed networks** throughout. It is imperative that these safety critical systems of systems are supported appropriately. Furthermore, for improved

teamwork in complex safety critical systems, it is imperative to define an appropriate STS model throughout the design lifecycle (i.e. iterative design of technology, procedures, training, implementation and evaluation) [51].

Human Oversight – The development of a trustworthy institutional design of human oversight is essential ([52], as EC Proposals dictate [53]), the oversight of AI should be done by "natural persons". As humans increasingly become "overseers" or "sense-checkers" and managers of operations (as opposed to fully active operators), it is imperative that they are fully and appropriately supported on all levels (i.e. Human-in-the-loop, Human-on-the-loop, Human-in-command).

Redundancy -Technology design frequently uses the human as main form of redundancy. A whole socio-technical system design needs to take this into consideration to ensure the humans in the system can cope, not overloaded, not out of the loop, trained adequately to cope for being main redundancy factor. This is linked directly to quantification of uncertainty with human as main redundancy factor for new technology designs using AI, Remote operations, Distributed teams, human-machine teamwork.

Scenario-based modelling (at strategic, operational and tactical levels) for such redundancy conditions is required to ascertain suitability for new procedures, communication and technology design using AI including but not limited to:

- Normal operations: fatigue, boredom, rest periods & breaks
- Non-normal operations: Incapacitation, Technology failure, Comms link failure, Cyber-Security threat

Neither the human agents nor AI are intended to work in silos, therefore an **integrated view of the human and non-human agents** and their contribution to the overall operation/mission/system is required ([54], [55], [56]). There is a continued lack of full understanding of the distribution of tasks between human and non-human agents in teams and the implications for operational safety are far reaching. The state of the art in STS modelling has yet to take full account of new teamworking and collaboration using AI [48].

There is a wealth of research done on **distributed situational awareness** ([57], [58], [59]) whereby the importance of recognizing situational awareness at a systems level and not individual or team level has been amply demonstrated [48]. STS level approach to situational awareness should be a key requirement for further research in this area with respect to AI. Such approaches should consider all stakeholders in the STS. The importance of learning and training in teams cannot be understated as well as team situational awareness, workload management and team co-ordination [60].

Technical Robustness and Safety -Technical capability and functionality from advances in AI cannot be denied, however, do such advances necessarily make the global sociotechnical system (which includes human stakeholders) safer or any more effective? Does increased speed of decision-making using AI naturally result in the best decisions being made? Does this equate to robust design or safe operations? Previously termed "human error assessment" should now be considered at a systems level – indeed more aptly named "system error assessment" would be in-keeping with state-of-the-art modelling of Socio-technical Systems [48]. Prospective Risk analyses would benefit future defence activities ([61], [62]) with a view to improved whole STS design.

Data Governance- all use of AI, monitoring and data usage / storage should be aligned with a clear data governance framework – i.e. aligned with a clear and transparent data justice perspective. It must be made apparent to all human agents that their data will not be used for any purpose other than the greater good and will be treated as part of a just culture. Mindful

governance models for managing information and strategic risk management are also welcome [63].

To address these data and AI challenges in the defence sector, a comprehensive Data and AI Governance Framework is essential. This framework integrates people, processes and technology and establishes the policies, procedures, standards, regulations, and tools, necessary for effectively managing an organization's data assets. By actively engaging stakeholders at all levels, this framework promotes a culture of accountability, continuous improvement, and responsible AI use. It ensures AI systems are developed and deployed in accordance with Trustworthy AI.

This framework encompasses several key components. First, the definition of roles and responsibilities ensures accountability within the organization. Second, data modelling, aiding in understanding relationships among different domains, entities, and attributes. This structured approach not only enhances understanding but also supports interoperability, allowing different systems to communicate and share data seamlessly.

Data lineage and traceability are also critical, as they track origins and transformations of data, ensuring data quality and integrity. Additionally, successful metadata management enhances usability by contextualizing data. This includes creating Data dictionaries and glossaries to standardize terminology, crucial for ensuring interoperability among NATO Allies. Moreover, implementing data catalogs helps manage data assets and facilitate discovery.

Furthermore, these resources will help classify privacy and security requirements for attributes, enabling effective access and sharing of information. All together, these elements play a vital role in supporting data and AI governance, risk management, and data security and privacy.

To sum up, a strong Data and AI governance strategy is essential for aligning with business objectives, driving value and fostering a culture of data stewardship across the organization. This approach enables strategic, tactical and operational efficiency, facilitating better-informed decision-making, and enhancing trust in data, all while ensuring compliance with regulations.

Transparency and Traceability – The use of AI and ethical practice surrounding defence operations should be made transparent and traceable. All human agents must be made aware of how their data will be used and how they will be protected for further progression within organizations, future recruitment etc. Furthermore, Transparency, Accountability, Justiciability, and Legitimacy are considered essential attributes in the prevention of institutionalised distrust in AI systems [52]. Further research in this area is warranted specific to defence and human-machine teamwork.

Accountability- Clarity of who is accountable at every stage of an operational process is fundamental especially when dealing with safety critical / life critical decisions. Such clarity needs to be made explicit on international, national, organizational, team and individual levels. A full modelling of the STS is required for this as is need for clarity on accountability made a key requirement.

Significant investment will continue to be made in AI technology design. The modelling of systems and requirements analyses of said systems should not solely be focussed on the procedural aspects of what the AI does, but how it impacts the human agent in terms of communication, collaboration, decision making, situational awareness and trust. Defence operations can only be made more adaptable and effective if all agents (i.e. human and non-human) are supported appropriately throughout the entirety of operations.

The significant investment in technology design should be matched **by significant investment in training, research, and thorough human factors evaluation** of how all team

members are supported for both normal and non-normal operations. Further implementation of AI will also have substantial impact on competence of organizations, i.e. training, selection and recruitment must adapt to new demands and skills bases required [64]. **Strategic Human Resource Management** for the foreseeable future should be cognisant of this. Regulatory and Legal support for training, selection, certification, and simulation must be put in place to support all activities and operations for EDA and EU MS Forces using AI.

Explainable AI Research on eXplainable AI (XAI) is leading the way in developing methods and techniques aimed at helping users understand unpredictable and unobservable AI systems. Such unpredictability may come from, for example, edge-cases in which AI is unable to handle a scenario because it was not anticipated during development. When a user is unaware of those edge-cases it will have a negative effect on the trustworthiness of the AI system. These edge-cases become even more impactful for autonomous and complex systems that are increasingly difficult to comprehend. As such, there is a need for AI systems to become explainable and understandable to humans without overburdening the user with information. Research into XAI is providing guidance on how to structure AI systems' interfaces to deliver timely explanations tailored to the user's requirements and context of use.

Exposing Bias Complex algorithms are used by AI to perform tasks that humans are not capable of. Despite obvious advantages to using AI, one substantial risk is that AI can produce algorithms that can cause a disadvantage for certain groups in society. This may cause further prejudice, discrimination, and marginalisation in society [65]. The EU Agency for Fundamental Rights (FRA) highlight that a “rights-based” approach is necessary to expose and prevent bias. In addition to this, some guidelines are available to avoid unfair bias [35] to question the quality of data [66] provide a solid starting point. This is critical, especially as there is a dearth of agreed standards for data quality assessments for machine learning applications. Whilst these questions and guidelines are welcomed, it is important that they are applied for every stage and iteration of the design lifecycle. FCAS forum commenced the specialization of ALTAI for defence [32]. This is a positive step in the right direction as it is a multi-disciplinary approach deploying use-cases. This work needs to be progressed beyond rules of engagement and extended to the whole lifecycle.

Dynamic Degrees of Autonomy: Traditional perspectives on Human-Machine Teaming typically classify distinct operational modes, such as the levels of autonomy, that are generally fixed for a particular Human-AI interactions. To truly realize the potential of Human-Machine Teaming, it is essential to develop AI technologies that are more adaptable and responsive to the user's current state and requirements. This flexibility in AI systems should mirror the fluid and responsive nature of human interactions within successful human teams. By adopting such an approach, AI can more effectively complement human abilities, leading to more intuitive and productive collaboration between humans and machines. This dynamic could enhance decision-making processes, improve the efficiency of operations, and lead to innovations in various fields where collaborative intelligence is crucial.

Developing and Evaluating Methods for Validation and Verification of Human-Machine Teams

Modelling human-AI interactions effectively presents significant challenges due to the inherent complexity and variability of human behaviour, as well as the dynamic nature of team tasks and structures. It is not feasible to model every single aspect of a system, however, the relationship between the system levels and how they impact the stakeholders in the system has to be interrogated from an integrated perspective. The process of evaluating these teams' collective capabilities and ensuring MHC involves a comprehensive, interdisciplinary approach that considers every facet of the Human-Machine Teams (HMT), including the human participants, the machines involved, and their interactions.

The conventional method of building effective HMTs involves assessing the individual capabilities of humans and AI systems, then finding optimal ways to integrate them to allow for safe, effective, and efficient integration of AI in HMTs for military applications. Of particular importance is defining the correct criteria to evaluate HMTs and how they can be measured.

Examples of these are:

- Observability. The ability for the human and AI-system to understand and comprehend each other (e.g., explainability, traceable behaviour and situational awareness).
- Predictability. The ability for the human and the AI-system to predict uc 14(to a degree) each other's behaviour (trust, intent recognition, and what-if reasoning).
- Directability. The ability for the human and the AI-system to direct each other's behaviour (E.g., delegation, work agreements, dynamic task allocation).

In addition to the qualitative benchmarks that remain pivotal during the developmental and evaluative stages, there is a need to include predictive computational models. The absence of such models would likely make the ongoing evaluation of HMTs impractical, given the extensive time and effort required to test and certify them across diverse operational contexts. These models must not only assess performance but also evaluate ethical considerations, team dynamics, interfacing capabilities, and other relevant factors.

To address these challenges, innovative engineering methods are required that integrate traditional verification and validation (V&V) techniques, such as model checking, simulation-based testing, and user validation through experimental trials.

Socio-technical System Risk: Further to research advising that human error be viewed as socio-technical systems error ([48], [51]). It is prudent to ensure that there is adequate feedback of risk information for the purposes of evaluation, system improvement, effective change management, monitoring of operational risk, performance, safety and other KPIs. Such KPIs should include ethics not only for auditing and monitoring purposes, but also for data governance and fairness around data gathering on human operators and other human stakeholders. The feedback should be appropriately linked to reporting system(s) to ensure that there is optimum use of data (i.e. transparency, ethical practice, effective use and resource management). From a systems perspective, it is critical that there is an integrated approach to risk management as opposed to consideration of risk from an individual technology level. The interaction of technologies and systems with all stakeholders across strategic, operational and tactical levels is essential if a rich picture of risk is to be captured. This is necessary for proactive risk management to facilitate smaller, more incremental adjustments to a system rather than a reactive approach which can result in greater expenditure of resources and manoeuvres much later in time and are far from ideal in such time-critical and safety-critical environments as military operations.

6.4. Human Factors Requirements

The following requirements are an outcome of the TAID working group. These were developed through discussions with expert members. This group was multi-disciplinary in nature with members spanning ethics, law, engineering, defence force and human factors communities. This was an iterative process over a period of months. Whilst not a formal requirements definition process, consensus was reached that human factors requirements are necessary for the design, procurement, deployment, use, training, evaluation, certification, management, maintenance of any AI technologies in a socio-technical system. This is considered ever more critical when the AI is integrated in safety-critical systems for defence with potentially far-reaching impact on human life.

The requirements shown in Table 2 reflect the output of discussions from the working group on how AI should be developed to ensure it is trustworthy and that both the direct and indirect stakeholders trust in the wider system in which they operate.

Table 2 - Human Factors Requirements

Requirement ID	Requirement Description
HF-01	All system design and use will be done in accordance with the Ethical Requirements stated in Chapter 0.
HF-02	A Socio-technical Systems based approach should be used throughout the entire design lifecycle.
HF-03	Detailed Socio-technical systems modelling should be a fundamental part of AI system design including those with non-human team members.
HF-04	All operations deploying AI systems should be carried out under the umbrella of a just culture. All use of AI monitoring and data usage / storage should be aligned with other Data Governance Frameworks such as EU and made specific to the military domain. All use of data should be treated in accordance with a just culture. Defence Forces should look to other safety critical industries to see how they have benefitted from just culture. Future research on how this can be adopted across military domains is welcome.
HF-05	STS modelling should be used to define clear delineation of stakeholder roles for operations for all levels of autonomy (i.e. 0 to 5) and transitions between those levels. This is especially critical for human-machine teaming and decision-support.
HF-06	All systems should be transparent to facilitate meaningful human control – this includes training, procedures and rules of engagement to ensure that human stakeholders are aware of and can anticipate intent (where possible) and the likely behaviour of the system. This is especially important for recovery of action, prioritization of task recovery etc.
HF-07	Data Governance Frameworks should be supported by a robust and comprehensive governance structure consisting of people, process, and technology implemented through a socio-technical systems analysis methodology, and linked to relevant data, analyses, risk assessment, strategic risk management, certification, standards, and compliant with regulations.
HF-08	HMIs should provide clear and unambiguous feedback to the human stakeholders so that they are fully aware of who has control throughout the task/ operation/ mission.
HF-09	Strategic Resource Management should be deployed across all operational domains with a view to short, medium, and longer-term system design, acquisition, implementation, maintenance, and sustainability.
HF-10	Recruitment, Selection and Assessment Methods for Strategic Human Resource Management should be reviewed and updated. This should be carried out as part of the socio-technical systems modelling and analyses.
HF-11	Training and assessment methods for both individuals and teams should be reviewed regularly in alignment with socio-technical systems modelling and analyses. This is critical for human-machine teaming.
HF-12	Scenario-based modelling for conditions where the human stakeholder is the main form of redundancy should be considered. These should include but not be limited to 1) normal operations: fatigue, boredom, rest periods and breaks, 2) non-normal operations: incapacitation, connection loss, technology failure, comms link failure, cyber-security threat etc.
HF-13	Training of the human operator should be carried out (both individually and in teams) with sufficient frequency to ensure that the human stakeholders are able to maintain operational performance and competence – especially when human-machine teaming and human reliance on enhanced technological systems.

Requirement ID	Requirement Description
HF-14	Metrics for assessing Crew Resource Management should be reviewed to include novel objective means of measurement and new roles, procedures for operations and for threat and error management.
HF-15	Verification and Validation of AI systems should follow a risk-based approach, should include stakeholders from the outset and for the entire duration of the design lifecycle.
HF-16	Prospective Risk analyses should be applied to benefit future defence activities with a view to improved whole system design.
HF-17	Respective Stakeholders at the organizational, national/ regional, and industry levels should ensure that a systems approach to the design, implementation, Validation and Verification of AI systems is accomplished by a capable and multidisciplinary team. This team should include qualified Human Factors professionals and/or individuals with appropriate training, experience, knowledge, and expertise. These individuals should work closely with the Value Lead (See Chapter 0 Ethics Requirement E04)
HF-18	All AI systems should provide observability, predictability, explainability, and directability for all levels of the STS.

7. Ethical Concerns Surrounding Trustworthiness for AI in Defence

7.1. Introduction

There are several challenges in ethical practice of which understanding the ethical concerns surrounding trustworthiness for AI in Defence and the law are merely the beginning. For stakeholders (i.e. those who interact with any technology involving AI or their integrated systems), it is important that they understand how ethics is relevant to their job, the tasks they have to perform and how their actions are likely to impact the wider system around them. This is critical for every stakeholder, from the commander in the field, those under their command, the technology developers, trainers, those responsible for procurement etc. The translation of abstract concepts into common language, terminology and guidance is fundamental for a diverse range of stakeholders to understand. Furthermore, in designing future technologies and systems there are several challenges that must be overcome to ensure ethical design and practice for the use of AI. For example:

1. Identifying **key stakeholders** who need to be involved in the AI system's design lifecycle.
2. Recognition of the **values** and laws relevant to the design process and translating them into requirements for new design.
3. Identifying how a new design **impacts** other areas of the system and the manner in which they are designed and function, such as:
    a. Other technological functions within wider integrated systems
    b. Procedural Design
    c. Non-technical skills such as communication and interaction between human and non-human team members
    d. Competence (individual competence, team competence)
    e. Training (including Training Needs Analysis, Training Design, Training Implementation and Evaluation)
    f. Strategic Risk Management
    g. Strategic Human Resource Management
4. **Measuring** system behaviour with and without human delegation ensures responsible AI use. Metrics for transparency, accountability, and fairness help assess ethical performance under varying levels of human control.
5. **Certification** frameworks validate AI systems' adherence to ethical and legal standards. This includes technical safety, data protection compliance, and unbiased, reliable system performance.

This chapter highlights a number of these challenges and identifies steps that can be taken to address them.

7.1.1. Key ethical requirements for Trustworthy

The EU produced a set of requirements for ethical design, all of which will be adopted by the European Defence Agency (EDA). The Ethics Guidelines for Trustworthy AI [23] offers practical guidance for achieving and implementing Trustworthy AI. It identifies a set of seven

key requirements that should be met, and which are based on the principles established at [23] Chapter I (Lawful, Ethical and Robust). The seven key requirements are discussed in further detail in Section 6.3.

The Seven Key Ethical Requirements for Trustworthy AI are as follows:

1. **Human agency and oversight:** Including fundamental rights, human agency, and human oversight.
2. **Technical Robustness and Safety:** Including resilience to attack and security, fall back plan and general safety, accuracy, reliability, and reproducibility.
3. **Privacy and Data Governance:** Including respect for privacy, quality and integrity of data, and access to data.
4. **Transparency:** Including traceability, explainability and communication.
5. **Diversity, non-discrimination, and fairness:** Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.
6. **Societal and environmental wellbeing:** Including sustainability and environmental friendliness, social impact, society, and democracy.
7. **Accountability:** Including auditability, minimisation and reporting of negative impact, trade-offs, and redress.

Note that the interrelationship of the seven requirements is of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle. Additionally, these requirements apply to all stakeholders involved in the life-cycle of AI systems, including developers, deployers, end-users, and the wider society.

7.1.2. Role of Ethics in AI Trustworthiness

Ethics is an academic discipline which is a subfield of philosophy. In general terms, it deals with questions like "What is a good action?", "What is the value of a human life?", "What is justice?", or "What is the good life?". In academic ethics, there are four major fields of research: (i) Meta-ethics, mostly concerning the meaning and reference of normative sentence, and the question how their truth values can be determined (if they have any); (ii) normative ethics, the practical means of determining a moral course of action by examining the standards for right and wrong action and assigning a value to specific actions; (iii) descriptive ethics, which aims at an empirical investigation of people's moral behaviour and beliefs; and (iv) applied ethics, concerning what we are obligated (or permitted) to do in a specific (often historically new) situation or a particular domain of (often historically unprecedented) possibilities for action. [23]

Applied ethics deals with real-life situations, where decisions have to be made under time-pressure, and often limited rationality. AI ethics as a subset of technology ethics focuses on the ethical aspects of the development, dissemination and application of AI over the entire life cycle of an Autonomous and Intelligent System (AIS). It deals with issues such as the responsibility of developers and technology companies, the moral evaluation of technology decisions, the protection of values through and despite AI and its impact on different population groups. [23] The term "ethical" refers to an evaluative assessment of such concrete actions and behaviours from a systematic, academic perspective.

While trustworthiness in AI refers to attributes such as reliability, accuracy, robustness, and transparency, these qualities alone do not guarantee that an AI system is ethical. A trustworthy AI system can perform its tasks effectively and predictably, gaining user confidence, yet its actions or decisions may conflict with ethical principles, such as fairness or justice. For example, a highly reliable credit-scoring AI may perpetuate systemic biases against certain demographics, highlighting that trustworthiness and ethics, while related, are distinct

concepts. Understanding this distinction is crucial for ensuring that AI systems not only inspire trust but also align with broader ethical considerations.

7.1.3. Problematic value lists

Since 2016/17, supranational organizations such as the United Nations, NATO and the EU, as well as national governments, non-governmental organizations and companies have been addressing the topic of "ethical AI". They have published several hundred AI strategy papers, around 80 of which deal explicitly with social concerns and compliance with the legal and ethical principles of AI.

Those 80 or so papers contain lists of attributes that AI systems should have[4]. The lists converge on largely the same attributes, including explainability, safety and non-bias. They are intended to serve as guidelines for AI developers to ensure that their AI systems meet minimum ethical standards. But they do not live up to this claim:

- The list of system attributes is incomplete and not necessarily relevant for Defence AI. For defenders, factors like effectiveness, proportionality, self-protection or peacekeeping play a role—but none of these values are included in the available lists.
- The lists are abstract and general. They give no indication of how to implement value protection in an AI system. Developers must therefore consider how to make values tangible and measurable. Lists and even detailed taxonomies do not provide any guidance for this either.
- In postmodern times, people talk a lot about values, ethics and morals, but without any sound (theoretical) knowledge about them, without context, without ontology, without knowledge of value theories or philosophical and historical contexts. Even enumerations of AI attributes often only list common quality attributes of (software) systems, but do not help to objectify and conceptualize values in such a way that developers are able to actually derive concrete system functions from such lists.
- In some cases, values can come into conflict with each other. For example, the pursuit of freedom can conflict with the pursuit of security. Developers have to weigh up and decide how to deal with such conflicts. Lists of quality attributes do not help in resolving conflicts of values.

Overall, lists of quality attributes do provide guidance as to what an organization or nation considers ethically, socially, and legally relevant for an AI system. However, they quickly reach their limits. What is needed, therefore, is an approach that allows us to go from the general to the detailed and translate general requirements into technical functions.

7.1.4. Value-Based Engineering

With the rise of Autonomous and Intelligent System (AIS), **VALUE-BASED ENGINEERING** (VBE) [67] has emerged in the system development ecosystem. Value-Based Engineering may be defined as: "Value-based Engineering" is a corporate innovation and engineering practice that caters to this new value-centred thinking. It supports a new era, where IT systems—the value bearers—are built to contribute to society's flourishing, while prohibiting negative side effects. The implication is that systems are not created any more merely because they maximize profit, are somehow useful, or embed new technical functionality; instead, technology's outspoken role is to support what is good, true, beautiful, peaceful and worthy in life. With this mission,

[4] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. external page: https://doi.org/10.1038/s42256-019-0088-2

Value-based Engineering seeks to extend non-functional requirements engineering, usability- or UX efforts. It is also different from technology impact assessment approaches in that it does not focus exclusively on the risks of value harms. In contrast, it strives to build “technology for humanity.” [67].

VBE comes into play early in the life cycle of an AIS—as management of **ETHICAL VALUE REQUIREMENTS** already in the design phase and even before the development of an AIS—and is therefore familiar territory for product/project managers and system developers. Requirements management that focuses on non-functional aspects of an AIS can enable the vision of the “good” early on in the life cycle of an AIS and before its implementation.

In contrast to the OECD, NATO or EU quality attributes lists, VBE does not work with lists of values or principles at all. Instead, VBE encourages developers to actively think themselves and non-biased about values—together with other stakeholders of an AIS. To allow for this, VBE first prompts developers to contextualize and detail an AIS that is intended to realize certain values. Context, concept of operations (ConOps) and system boundaries of an AIS are a crucial first step in identifying values that are affected by an AIS both positively and negatively.

7.1.5. Value-based Engineering at a glance

7.1.5.1 The process

The concept of operations is the starting point for Valued-Based Engineering (VBE). ConOps is a document in simple user language that defines the basic concepts and functions of the AIS under investigation.

With ConOps, VBE requires a description of the Autonomous and Intelligent Systems (AIS), its context, the political, ecological, organizational and social circumstances of the system as well as its components, interfaces, data flows, stakeholders and system boundaries. Contextualization characterizes this operational concept, and its result is the **SYSTEM OF INTEREST**. The advantage: instead of “the AI” or AI in general, the focus shifts to concrete goals and functions of a definable and delimited AIS.

To create a ConOps document, VBE encourages the stakeholders of an AIS to imagine what impact it would have if the AIS in question were rolled out globally. The motto is: Think Big! Imagine what the benefits and disadvantages would be if the AIS in question was rolled out at a global scale and affected not just a few, but millions of stakeholders. With such conceptual scaling, the AIS will cover a very large universe of values. ConOps therefore lays the foundation for value exploration. The more stakeholders are affected by a system, the larger the universe of values that a system must respect.

Only when ConOps is available can an open exploration of values take place. This involves describing the ethical, legal and social effects of the AIS, assigning values to these effects and conceptualizing and prioritizing the values themselves. It is not until this ethical design of the AIS takes place that the Ethical Value Requirements are defined, in order to ensure that the system has no negative impact on society or the environment. As work results and conformity requirements of the AIS, they differ significantly from the stereotypical lists that are repeatedly produced everywhere.

At the end of the process, a risk-based system design is carried out and measures are taken to mitigate the greatest risks that ethical value requirements may not be met.

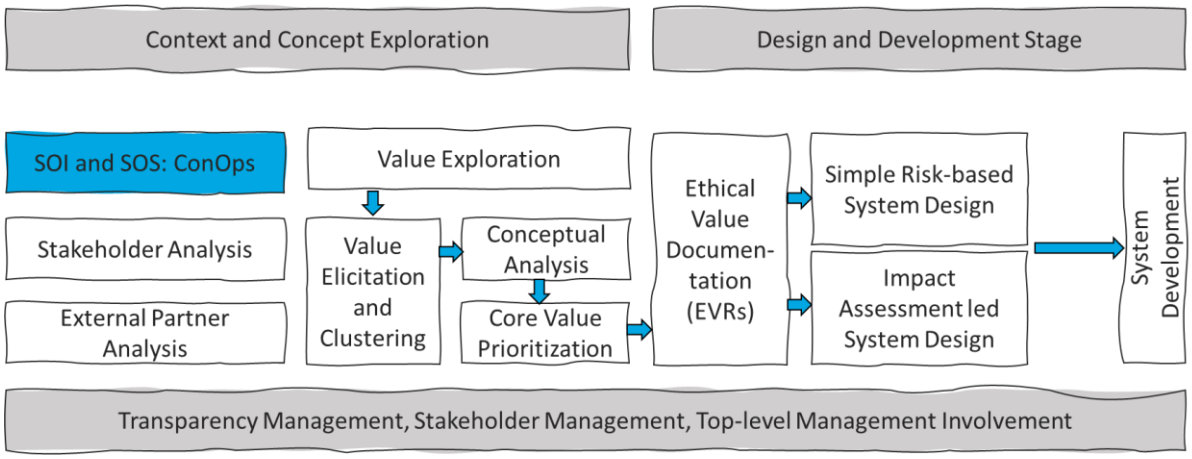


Figure 12 - The process of value-based development in the early life cycle of systems

As illustrated in Figure 12 the process is so fundamental that it allows abstraction from the technical system as well as the development of value-based business models, for example for the operation of AI systems.

Stakeholder Definitions for Value-based Engineering

Stakeholder: An individual, organization, group, or other entity that can affect, be affected or perceives itself to be affected by a System of Interest. Stakeholders have a legitimate right, share, claim, influence or interest in a system or in its possession of characteristics that meet their needs, expectations or are relevant for their sustainability. As highlighted in Chapter 2, stakeholders can be of direct or indirect nature. A stakeholder advocate may also be appointed. An advocate is a steward who is entrusted with the responsibility to take care of representing the interests of those stakeholders who cannot participate in activities or the application of standards themselves. He or she should have the credentials to speak for the interest group, person or entity s(he) represents, i.e. they *also give a voice to all those entities who are passively affected by the system or might be so in the future (including nature, animals, unborn, minorities).*

7.1.6. Value-based Engineering with ISO/IEC/IEEE 24748-7000:2022

To ensure the entire design lifecycle is ethical and compliant with values, i.e. any relevant value, even if not listed on any list, ISO/IEC/IEEE 24748-7000:2022 [68] provides a process:

- To prioritize values as appropriate,
- To make ethics measurable,
- To translate ethical requirements into system features,
- All based on sound philosophical grounds,
- With a suggestion for AI readiness of an organization,
- With a suggestion for “ethical conformance”.

This whitepaper suggests that the EDA follows a norm-based approach for Defence AI standardization and observe other existing standards. Note that ISO/IEC/IEEE 24748-7000:2022 [68] is the sole standard globally which combines philosophy and engineering, providing a structured framework to integrate ethical values into the engineering processes ensuring Defence AI systems are ethically compliant across the lifecycle.

7.1.6.1 A new profession: the Value Lead

Value leads drive a system's value mission. It is hard for developers to venture into philosophical thinking. The conception, design and implementation of value-based Autonomous and Intelligent Systems (AIS) requires people of integrity who not only have knowledge of values, but also practice ethical thinking on a daily basis and throughout their lives. To ensure such a value orientation, VBE is introducing the job profile of the Value Lead. This is a "person tasked with coordinating and carrying out tasks related to the identification and prioritization of ethical values and the traceability of values from requirements and design artifacts" [68]. Such a person is therefore knowledgeable about values, value theories, ethical theories and philosophy, but at the same time technically proficient. Value Leads are part of the system engineering team, helping them determine value and later monitoring and documenting that the system design is ethically consistent with the values found relevant to the system. A value lead can be considered as a stakeholder advocate.

7.1.6.2 From values to Ethical Value Requirements

Once the stakeholders have defined, conceptualized and prioritized values, VBE requires the definition of relevant ethical value requirements (EVR) for the system. EVR represents the ethical, social and legal needs and expectations of the stakeholders of an AIS. They are market requirements, as opposed to system requirements. Starting from prioritized core values, each value quality is translated into an EVR and into either (1) a programmable system function or (2) an organizational action, each with the aim of realizing a specific value quality.

7.1.6.3 What risks are EVRs exposed to?

As part of a risk-based design, the stakeholders of an AIS define which risks EVRs are exposed to. This risk-based design approach ensures that organizational measures to preserve value are implemented immediately and without hesitation or further analysis. Low technical risks can be assessed using a risk matrix; risk mitigation measures are then treated and prioritized like any other functional system requirement. For example, in an agile system development process, they are backlogged.

However, if core intrinsic values are threatened, this is always classified as high risk. The risk assessment should then be followed by a technology assessment.

7.1.6.4 VBE transparency requirements and verification?

Value-Based Engineering (VBE) was standardized for the first time in 2021 in IEEE 7000™ (today: ISO/IEC/IEEE 24748-7000:2022 [68]). A system's compliance with VBE requirements can, therefore, be verified as follows:

- Either completely, i.e. the VBE process was followed, and the necessary compliance forms were provided;
- Or partially, by demonstrating either process compliance or results. The latter include a system description, curated lists of values based on three ethical theories, value clusters and prioritization of high core values, EVR and a risk assessment.

If a person has been trained as a Value Lead, they too can be accredited and certified, and not just an AIS. However, the knowledge that a Value Lead acquires during their training is not enough to become certified. A Value Lead must demonstrate personal integrity and moral behavior. This requires a lifetime of practice. Value leads are therefore only accredited by the IEEE for a certain period; in the event of misconduct, their accreditation can be withdrawn.

7.1.7. Open Issues

This whitepaper serves to highlight some of the open issues surrounding the Trustworthiness of AI for Defence. Meaningful Human Control is one of the key issues.

Value-Based Engineering and a socio-technical systems approach are of equal importance. There should be due consideration for value sensitive design from the outset of the design lifecycle.

ALTAI for Defence: Given the criticality of ethics to both Defence Operations and the use AI, it is imperative that there is a common means of reference and guidelines for all operations throughout the design lifecycle. ALTAI [35] and FCAS [32] have made commendable in-roads, however, a defence-specific version in which both systems and value-based engineering approaches is required. Self-assessment tools alone do not provide sufficient guidelines for stakeholders to know the difference of what change needs to be implemented for them to be able to score a 4/5 instead of a 2/5. Context such as the FCAS working group[5] provide use-cases, but stakeholders need to understand how to make guidelines work for their own operational domain (OD). This is far more likely to be achieved through understanding the values and requirements around the ethical principles as lead by an accredited Value Lead. Value leads can be viewed as the new “leaders” in this field for effecting change and driving ethical design & practice. This is not likely to be a quick process and will be one involving a change in culture to increase Trust and Trustworthiness of AI for defence.

Meaningful Human Control (MHC) is one of the most contentious areas concerning Trustworthiness for AI in Defence (TAID). Particularly in reassuring the public, the EDA will consider additional research going forward in this area. Recent Work [69] highlights the complex nature of MHC and the contributions that Compliance, Dignity and Responsibility make to the debate. Eggert, L. also cautions against considering MHC to be a solution to the complex moral and operational practicalities of AI in defence- in particular, for Autonomous Weapons. Further exploration of these issues can be found in the case study in the appendices.

Human in the loop AI integration into defence systems poses unique ethical challenges. The concept of human-in-the-loop ethics is crucial to ensure that AI applications in defence adhere to moral, legal, and ethical standards while enhancing military capabilities.

Human Operator Definition: The human operator, in the context of AI in defence, refers to the individual responsible for overseeing, controlling, and making critical decisions within AI-enabled systems. Their role is pivotal in ensuring that ethical considerations are upheld throughout the use of AI technologies.

Responsibilities for Human Operator: Human operators bear the responsibility of maintaining accountability and oversight in AI-driven defence systems. They must possess the competence to interpret AI-generated recommendations, intervene when necessary, and mitigate potential risks or ethical violations. Furthermore, human operators are tasked with upholding humanitarian principles and adhering to international laws governing the conduct of warfare.

System Requirements for Ethical AI: Ethical AI systems in defence must be designed with transparency, accountability, and fairness in mind. They should prioritize the explainability of AI-generated decisions, enabling human operators to understand the rationale behind algorithmic outputs. Additionally, AI systems must be regularly audited to detect biases, errors, or unintended consequences, with mechanisms in place for swift intervention and correction.

[5] https://www.fcas-forum.eu/en/

No Human-in-the-Loop Cases: While human-in-the-loop ethics is essential for ensuring the responsible use of AI in defence, there may be scenarios where autonomous AI systems are preferred or necessary. These cases typically involve time-sensitive operations or environments where human intervention is impractical or poses greater risks. However, even in such instances, rigorous ethical guidelines and fail-safe mechanisms must be implemented to minimize the potential for unethical outcomes.

In conclusion, human-in-the-loop ethics serves as a critical framework for guiding the integration of AI in defence. By empowering human operators with the knowledge, skills, and tools to oversee AI systems effectively, we can uphold ethical standards, mitigate risks, and ensure that AI technologies align with moral imperatives and societal values. Through transparent design, accountability mechanisms, and adherence to international norms, we can harness the potential of AI in defence while safeguarding against unethical practices and unintended consequences.

Full recognition is given to the dearth in appropriate Key Performance Indicators (KPIs) for assessment and measurement of Ethical practice and design for Trustworthy AI. Testing and Evaluation methodology framework for the Ethical requirements are essential and can only be done reliably through a systems approach linking each stage of the design lifecycle with relevant feedback and data. Whilst this will not be a simple journey to undertake, the EDA views this as one worthy of commencing and values further research in this area.

7.1.8. Recommendations

In conclusion, the following recommendations are made:

Table 3 - Ethical Recommendations

Recommendations	Description
E01	All design, use and testing of systems (representing the full design lifecycle) shall be done in accordance with Article 2 of EU Treaty, the Geneva Convention, the Ottawa Treaty, European Convention on Human Rights, the European Social Charter, the Charter of Fundamental Rights, and International Human Rights Law.
E02	All design, use and testing of systems (representing the full design lifecycle) shall be done in accordance with the 7 key ethical requirements as outlined by the EU High Level Expert Group and as adopted by the EU AI Act (2023) [70].
E03	All systems should be designed using a Value-Based Engineering Approach and the entire design lifecycle should be overseen by a named and certified Value Lead.
E04	A full definition of relevant ethical value requirements for the system should be carried out with a multi-disciplinary team of stakeholders under the supervision of the value lead.
E05	AI systems should support human agency and human decision-making, as prescribed by the principle of respect for human autonomy.
E06	All systems should prevent harm to privacy (a fundamental right) also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.
E07	All systems should be transparent and encompass three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system.
E08	All systems should enable inclusion and diversity throughout the entire AI system's life cycle and should prevent inadvertent historic bias, incompleteness, and bad governance models. AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender,

Recommendations	Description
	abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance.
E09	The principle of accountability should be applied, and mechanisms be put in place to ensure responsibility for the development, deployment and/or use of AI systems.
E10	All values relevant to the systems should be identified and assessed by the (named and certified) Value Lead and include a multi-disciplinary stakeholder group in the identification and assessment process (throughout the entire design lifecycle (design, implementation, Validation, Verification, Periodic Review).

8. Impact Analysis on AI Use Cases in Defence

This chapter will provide a basis to assess both Opportunities and Risks regarding the integration of AI technology for different usages in defence. Along with a methodology for impact analysis, the chapter provides a list of ongoing Use Cases on which a preliminary impact assessment is conducted. However, notice that this chapter only contains the analysis of one Use Case for illustrative purposes. The full list of Use Cases analyses can be found in "TAID Annex 04 AI Use Cases for Defence". Overall, this exercise aims to provide guidance to decision makers by facilitating means for assessment.

8.1. Methodological Aspects

8.1.1. Purpose/Scope of the Analysis

The analysis in this chapter aims to breakdown the relationship between AI and defence, considering both **Opportunities** and **Risks** impacts, which will be presented and discussed in this document.

Understanding the Opportunities impacts of AI in defence is essential for strategic reasons. AI will facilitate modernisation in various aspects of the sector and increase of operational efficiency, as it acts as an enabler of autonomous systems, real-time data analysis and strategic decision making.

However, it is also necessary to identify and address potential Risks impacts, which may limit the use of AI techniques.

8.1.2. Adopting a Definition for Evaluation of Impact

The impact of AI technologies is to compare (with or without) if at an affordable cost AI in defence improves the operational efficiency or modernises some aspects of equipment and operations to meet the evolving needs of modern armed conflicts or to increase the risks for the society.

This definition relies on the following assumptions:

- A significant impact that should be recognized in future battle systems operating through the use of AI, is on cost. The cost parameter affects both operational cost and equipment cost. Given the fact that modern conflicts finally evolve in wars of attrition, the primary cost to be minimized is the one of human life. Secondary, cost of operation and equipment cost are critical factors to be considered. The use of AI should contribute to the increase of the efficiency of the equipment in terms of cost, i.e. (cost of friendly equipment used) < (cost of enemy equipment destroyed), otherwise the whole operation is illogical.
- When talking about equipment and operations, the system of interest is to be characterised:
    - The "Defence systems" which Defence sector delivers to the Armed forces, for them to operate.
    - The way Defence sector actors design, develop and maintain the "Defence systems".
- As numerous other initiatives also deal with AI trustworthiness, the subgroup has pointed out the need to concentrate on what is peculiar in Defence, both the specificities of the Defence systems, as well as the specific uses of AI technologies in Defence.

8.1.3. Methodology

The methodology proposed to identify AI impacts is summarized in the following items. This methodology may require the involvement of different multidisciplinary experts and stakeholders in the defence sector.

Step 1: Identify characteristics to measure impacts. The introduction of AI technology in defence shall have implications depending upon the specificities of the system, the environment of operation, and the missions defining system usage. However, impacts can go beyond referred aspects and reach other dimensions, for instance the Observe, Orient, Decide and Act (OODA) loop[6]. To assess Risks and Opportunities, commonly used Characteristics to measuring impacts should be identified. This process mostly relies upon expert knowledge and depend upon specific military sector practices and stakeholders. As a reference, in Table 4, some of the most common characteristics used to evaluate impacts are listed and categorized. Notice that this list is complementary to the list of properties already introduced in Table 1. Also notice that for the purposes of evaluating human factor aspects, the requirements introduced in Table 2 can also be used.

Step 2: Select appropriate characteristics. The outcome of this phase should provide the most appropriate characteristics to evaluate impacts, and among them, the selected subset to conduct the impact analysis. Some aspects to consider when addressing the selection are:

- **Prioritizing human factors.** It is imperative to identify high-level aspects useful to drive the impact analysis. For instance, human live and safety are usually the most important criteria whereas cost and operation come second in priority.
- **Prioritizing military specificities.** Prioritizing is suggested when a wide variety of characteristics may apply to evaluate impacts. In such a case, the traits specific to the defence sector, especially those not already covered by civil standards or practices, can be considered as priority.
- **Capabilities gain.** The gain in capabilities is a stake when selecting characteristics to evaluate impact, since they help to position augmented capabilities vs. opponents' capabilities. This is appropriate when the protections' efficiency is challenged by opponents' capabilities given an exposed surface and a window of opportunity.
- **Organization evolution and change management**. As a result of integrating AI into military systems, an evolution in the organization is foreseen, particularly when increased levels of autonomy are sought. This transition will require effective change management strategies to ensure smooth implementation. Organization adaptations may be required in the command chain not only at individual but also at collective level and may involve strategy, operational and field personnel. Change management will be crucial in guiding these adaptations and helping personnel at all levels adjust to new roles and responsibilities.
- **Policy and value alignment.** In most cases, AI systems shall comply with applicable policies, regulations, and standards. The rules and requirements therein shall be considered specially when dual usage of the AI system, in civil and military domains, is foreseen.

Step 3: Describe system usage. A Use Case description allows to detail not only the system, including the AI component/technology, but also the involved stakeholders (users, developers, authorities) who play a role during impact assessment. Detailing the system purposes

[6] OODA Loop: https://en.wikipedia.org/wiki/OODA_loop, https://doi.org/10.1016/j.ijme.2022.100703

(containing the AI/ML components), its context and category (military, civil, dual) provide further elements to properly assess Risks and Opportunities.

Step 4: Apply selected characteristics. The selected characteristics in step 2 are applied to evaluate the system usage specified in step 3. To do so, the analyst searches the distinctive traits of the system producing effects on the characteristics in question. The findings are listed and summarized according to positive and negative effects, then finally appreciated as advantages or drawbacks. Referred outcomes constitute the Opportunities and Risks categories.

Step 5: Acceptability of Opportunities vs Risks. An assessment is conducted to decide on the acceptability of Opportunities vs. Risks. The usage of AI systems should exhibit acceptable trade-off between Opportunities and Risks. When there is no perceptible gain, a Neutral impact is assumed.

The application of previous methodology is illustrated in Table 5 relying upon the impact Characteristics listed in Table 4 and also Table 1 to fictional scenarios involving instances of AI usages and stakeholders.

8.2. Main Characteristics for Impact Analysis

This subsection includes two items essential to understand the impact analysis. In Section 8.2.1, a list of Characteristics in Table 4 to evaluate impacts is provided. In Section 8.2.2, the Characteristics are used to determine impact on fictional scenarios involving AI systems, as outlined in Table 5, relying upon the methodology in Section 8.1.3.

8.2.1. Checklist of Characteristics for Impact Analyses

The Table 4Table 4 includes a list of most common characteristics proposed to evaluate impacts of AI usage in Defence. The list summarizes expert knowledge gathered by the EDA-TAID working group and is complementary to the list of properties already introduced in Table 1. Notice that elements from both tables can be used in complement to the Requirements introduced in Table 2 which were elicited from human-factors experts' knowledge. The characteristics hereinafter are rather specific to evaluate impact of AI usage according to the following categories:

- System Performance,
- Development Lifecycle,
- Advanced System Design Characteristics,
- Military Operations,
- Military Specifics,
- Human-World Values.

Also, notice that the characteristics are not independent, and their usage may require a detailed description of the AI system, involving refinements, other impact characteristics, and other technical properties not listed herein.

Table 4 - List of characteristics for the assessment of impacts of AI technology in the Defence sector

ID	Impact characteristic	Description of the Impact Characteristic (Opportunity form)
System Performance		
PER-01	Throughput/Speed	The workflow throughput is improved so that the rapidity of the mission or function accomplishment are increased.
PER-02	Novelty	The AI component/technology or its integration and usage within a system constitute a novelty regarding the state of the art.
PER-03	Function gain, extension	The usage of the AI component/technology produces a function gain or an extension of existing functions.

ID	Impact characteristic	Description of the Impact Characteristic (Opportunity form)
PER-04	Safety-Performance trade-off	An Increased performance of the AI technology ensures fulfilment of safety requirements during mission accomplishment allowing better AI system reaction to critical situations[7].
Development Cycle/Life Cycle		
DCLC-01	Quality of design, DevOps, V&V (Faster /lighter /better)	The introduced approach, method, or framework addresses an AI stake or leverages an AI design/development activity or industry practice to improve quality, reduce time, etc.
DCLC-02	Cost of design	The cost of the engineering process (in time, effort, money) followed during the design phase of the AI system or technology is decreased or remains acceptable.
DCLC-03	Production performance	A gain in the streamlining operations: production, decision making, logistics, etc.
DCLC-04	Homologation/certification	Improvement in any of the process phases required as preconditions to approve and release the system to operation.
DCLC-05	Cost of maintenance	A reduction in the complexity needed to maintain the system (update, repair, upgrade) in terms of time, effort, etc. and increased efficiency thus reducing costs.
Advanced System Design Characteristics		
ASDC-01	Human-AI synergies	Smarter and smoother interactions between human and AI produce suitable human-machine synergy (win-win strategy).
ASDC-02	Increased AI TRL	The usage or integration of AI technology helps to demonstrate or is a prerequisite for an increased TRLs (Technology Readiness Level).
Military Operations		
MOP-01	Risks acceptability	Increased levels of acceptability for risks and residual risks during system missions' assessment.
MOP-02	Military goals and objectives	The usage of AI allows to achieve new military goals and objectives.
MOP-03	Human lives preservation	Reduced human exposure to risks, including lower injuries severity and/or number of casualties.
MOP-04	Defence specific aspects	Improved level in any of the specifics of the Defence sector, for instance the ones related to the Observation-Orientation-Decision-Action (OODA) loop.
MOP-05	Cost of warfare	Reducing financial cost of conflicts through the usage of AI.
MOP-06	Human Resource Management	The management of military human resources is improved at any of its levels: strategy, operation, tactical.
MOP-07	Material Resource Management	The management of military material resources is improved at any of its levels: strategy, operation, tactical.
MOP-08	Mission accomplishment	The deployment of AI technology allows new missions accomplishment or increased mission scope.
MOP-09	Equipment Capacities	Improvement of capacities for onboard equipment (e.g. aircraft detection system) or on-ground equipment (e.g. soldier protection equipment).
Military Specifics (AI contribution to other properties)		
MSP-01	Attrition rates	Increased attrition rates in the military bodies: strengthen, robustness, resilience.
MSP-02	Opponent' surprise	Slowing down the effect of surprise coming from the opponent.
MSP-03	Surprise effect	Increased effect of surprise on the opponent.
MSP-04	Decision supremacy	Maintaining strategic superiority by ensuring that decisions within one's strategy outpace and outmaneuver the opponent's strategy. It involves dominating the combined Observation-Decision-Action (ODA) cycles by effectively observing, analyzing, and acting more decisively and accurately than the opponent. Decision supremacy ensures that strategic choices consistently counter or preempt the adversary's moves, securing a competitive advantage in dynamic and contested environments.
MSP-05	Armies' coordination	Leveraging mission impact by improved synergy during multiple armies' coordination.
MSP-06	Command intent	Improved accuracy, transmission, and secrecy of the own intent to intended parties.

[7] A detailed discussion to positioning Safety-critical vs. Mission-critical terminology can be found in the Appendixes.

ID	Impact characteristic	Description of the Impact Characteristic (Opportunity form)
MSP-07	Opponent's command intent	Improved assessment, estimation, disclosing of the command intent from the opponent.
MSP-08	Mission support	Improved mutual support and relief in critical warfare situations during missions.
Human-World Values		
HUV-01	Ethics	The deployment of AI technology fosters or ensures alignment with EU regulations in matter of ethics.
HUV-02	Human-AI race conditions	The usage of AI technology keeps the risk of human-AI race conditions acceptable.
HUV-03	Societal/economic stability	The usage of AI technology leads to societal or economic changes whereas their overall stability is still preserved

8.2.2. Illustrative Examples of the Impact Analysis Characteristics

Table 5 - Instances to illustrate evaluation of impacts on fictional scenarios (opportunities, risks, neutral).The Table 5 recalls in the 1st and 2nd columns, several of the impact characteristics already introduced in Table 1 and Table 4, which are latter applied to analysing impacts on a fictional scenario, and summarized in the 3rd column. The effects are finally classified in the 4th column according to the following categories: Risk, Opportunity, Neutral (see methodology in Section 8.1.3).

Table 5 - Instances to illustrate evaluation of impacts on fictional scenarios (opportunities, risks, neutral).

ID	Trustworthiness Property/Impact characteristic	Scenario Instance to Illustrate Impact	Impact Category
System Performance			
TAID-02	Accuracy	AI techniques are applied to traditional analytical methods in design phase affected by high uncertainty (e.g. structural loads evaluation in aircrafts) providing insights into their deficiencies and facilitating their enhancement.	Opportunities
PER-01	Throughput/Speed	Techniques are applied during data preprocessing to reduce computation time (e.g. filtering corrupted data) what leads to better system response time (without loss of accuracy).	Opportunities
TAID-21	Recognition	An AI system is used for obstacle detection, however whereas it offers augmented capabilities for recognition, it also exhibits non-negligible risk of false recognition.	Risks
PER-02	Novelty	A novel ML/DL approach already applied in the civil/research sector is applied into the military sector to replace an already automated function.	Neutral
PER-03	Function gain, extension	A component developed with traditional technology is replaced by an AI component without any function gain.	Risks
TAID-06	Autonomy	Increased level of system autonomy allows to perform unmanned warfare features.	Opportunities
TAID-37	Usability	An AI-based collision avoidance system certified according to civil aerospace standards is integrated in military planes without change of its operational design domain (ODD).	Opportunities/Neutral
TAID-27	Reusability	An AI-based component used for obstacle detection in road vehicles in the civil sector is re-used for military purposes but the lack of data for ODD re-design prevents re-usability.	Risks
PER-04	Safety-Performance trade-off	The performance of a ML component is increased to decrease failure rate during missions of a military aircraft over civil areas thus fulfilling a safety objective for certification.	Opportunities

ID	Trustworthiness Property/Impact characteristic	Scenario Instance to Illustrate Impact	Impact Category
TAID-05	HW capacity to support AI complexity	An AI-based solution requires a considerable amount of memory and computational HW resources preventing their adoption on small and embedded platforms.	Risks
Development Cycle/Life Cycle			
DCLC-01	Quality of design, DevOps, V&V (Faster /lighter /better)	Formal methods (e.g. surrogate models, abstract interpretation) are applied to implement rigorous data quality analysis during design phase, enhancing data value awareness and prompting a deeper understanding of outliers.	Opportunities
DCLC-02	Cost of design	An AI solution for which inference results are different than the trained model, requires a large amount of time to be validated due to duplication of efforts when analysing discrepancies.	Risks
DCLC-03	Production performance	AI technology is applied to optimize production and maintenance schedules, and to analyse large set of data for decision making thus reducing time and costs, however, the outcomes quality is decreased due to mishandled AI errors.	Opportunities/Risks
DCLC-04	Homologation/certification	A new AI system supports pilots during a critical phase flight, e.g. refuelling. The pilots need to pass costly training to ensure system acceptance/certification.	Risks
DCLC-05	Cost of maintenance	An AI algorithm is stable and robust enough w.r.t. its ODD so that the number of SW updates across component service lifetime is acceptable as compared to a typical SW item.	Opportunities
Advanced System Design Characteristics			
TAID-12	Explainability (AI trustworthiness attributes)	Explainability: DL techniques are applied to raw data for processing, extracting features automatically during development activity, however resulting in less explicable neural networks.	Risks
ASDC-01	Human-AI synergies	A visual-aid AI-based screen augments operator's visibility of the terrain ahead and the confidence in human decision whereas current operator's skills are still exercised.	Opportunities
TAID-04	AI self-protection	Components integrating AI algorithms are deployed with Physically Unclonable technology (e.g. PUFs) to prevent cloning by opponents (self-protection). However, reverse engineering can be applied to unveil the implementation.	Opportunities/ Risks
TAID-03	AI resilience	A defence system is integrated with security watchdogs, and safeguards to prevent AI to be cheated, diverted, or attacked.	Opportunities
ASDC-02	Increased AI TRL	A Use Case integrating AI/ML technology previously tested with experimental data is scaled up and tested in a real environment thus showing increased level of maturity.	Opportunities
Military Operations			
MOP-01	Risks acceptability	The integration of AI technology in military equipment reduces the need of human intervention or the exposed surface to opponents. The residual risk of formerly unbearable-risk missions becomes acceptable.	Opportunities
MOP-02	Military goals and objectives	The integration of AI technology fosters the achievement of certain organization objectives, formerly unachievable without such technology, but introduces uncertainty by the replacement of well-known, human-controlled technology.	Opportunities/Risks

ID	Trustworthiness Property/Impact characteristic	Scenario Instance to Illustrate Impact	Impact Category
MOP-03	Human lives preservation	The usage of AI technology to conduct reconnaissance of the battlefield helps to reduce human exposure to risks.	Opportunities
MOP-04	Defence specific aspects	AI technology helps to automate phases and tasks in the Observation-Orientation-Decision-Action loop. However, it unintendedly bypasses hierarchy principles in the command chain.	Risks
MOP-05	Warfare Cost	Usage of unmanned vehicles leveraged by AI technology reduces costs (e.g. training) without loss of confidence and boosting defensive capacities (deterrence augmented).	Opportunities
TAID-26	Responsibility	A new autonomous AI-based systems is deployed without proper re-assessment of roles and duties and the role and responsibility of some stakeholders become opaque.	Risks
MOP-06	Human Resource Management	An AI-based simulator based upon virtual reality is used for human training during critical missions what increases skills and survivability chances.	Opportunities
MOP-07	Material Resource Management	A face-recognition system is installed to control access to security-sensitive premises, the residual risk due to unintended access by third parties is reduced.	Opportunities
MOP-08	Mission accomplishment	The usage of unmanned aerial vehicle allows to intervene in extreme environmental conditions (temperature, pressure) thus increasing missions' scope.	Opportunities
MOP-09	Equipment Capacities	A night vision equipment integrating AI augments vision range and detection of moving objects thus helping to better identify risky situations.	Opportunities
Human-World Values			
HUV-01	Ethics	An expert is requested to conduct review to ensure that a new AI-based equipment provided by a subcontractor is aligned with EU AI Act and other legal and ethical precepts.	Opportunities
TAID-29	Sovereignty	An AI-based system demands bulky data collection from restricted defence zones. The databases and servers storing data and AI models are insufficiently protected to prevent data corruption, stealing/leakage, or loss.	Risks
TAID-32	Sustainability	An AI solution requiring a HPC platform (High Performance Computing) yields a large carbon footprint during operation without function gain, an inefficient energy usage is observed.	Risks
TAID-11	Controllability (incl. Meaning Human Control)	Some phases of a launch system are to be AI-automated. A risk and impact assessment are conducted to identify the functions that should remain under human control.	Opportunities
HUV-02	Human-AI race conditions	An AI system is deployed to automate a pilot-error prone task in a flight phase (e.g. approaching). The gain in automation makes the pilot loss essential skills to react and takeover: the pilot becomes unused to manoeuvre in a critical situation when AI deactivates (e.g. wind shear leads to AI system deactivation).	Risks
HUV-03	Societal/economic stability	The introduction of AI reduces exposure of human resources to warfare risks, the high degree of automation leads to replacement and re-adaptation of job positions via training.	Opportunities

NB. The adopted notation during impact analysis is the following: a Characteristic with code XXX-01 from Table 5 evaluated as a "Risk" is denoted by XXX-01:R whereas the same Characteristic evaluated as an "Opportunity" is denoted by XXX-01:O. Finally, if there is a neutral assessment, it is denoted by XXX-01:N.

8.3. Military Use Cases and Scenarios for Trustworthy AI

The objective of this section is twofold:

- first, to briefly introduce a set of ongoing military-related Use Cases for which AI technology integration is foreseen,
- second, apply the methodology described in Section 8.1.3 to illustrate the impact assessment on a selected Use Case.

The Table 6Table 6 includes a list of ongoing Use Cases integrating AI technology for Defence. There is no intent to be exhaustive, the purpose of the Use Cases is to both illustrate and challenge the impact assessment methodology. Notice that whereas only the impact analysis of UC01 is presented in following Section 8.3.1**Error! Reference source not found.**, t he complete list of Use Cases analyses can be found in "TAID Annex 04 AI Use Cases for Defence".

Table 6 - List of AI Use Cases for Defence

ID	Title	Actors	Systems of interest	Level
UC01	Decision-making in multi domain operations	C4I operators, AI subject matter experts	C4I system in multi domain operations	Tactical
UC02	Failure of a decision support system	Commander interacting people, Commander training team, authorities	Sensor Mesh (passive/active radar, EO sensors, audio sensors) to Optimize Situational Awareness	Tactical
UC03	Collision avoidance/swarming of drones/emergence (tactics to neutralise targets)	Remote pilots, pilots in the surroundings	Detect & Avoid system	Tactical
UC04	Mission training	Military commanders and other military personnel	Combat Training System	Operational
UC05	Aerial refuelling	aerial refuelling operator	aerial refuelling system	Tactical
UC06	Data-centric security	Both human and non-human annotators as well as cross-domain solutions (information processors/guards)	Security domains and information processors	Strategic
UC07	Military Approval/Certification	Manufacturers, Data providers, Air traffic controllers, Approval/Certification Authorities	Equipment being used for both military and civil applications	Strategic
UC08	Meaningful Human Control	Drone operator, System designer	targeting and decision making on Firing (Autonomous Weapon System)	Tactical
UC09	Active Autonomous Cyber Defence	System developers, Cyberoperator, system owner	A cyber security system	Operational

8.3.1. UC01 – Decision-Making in Multi-Domain Operations

Overall Description

Command, control, communications, computers, and intelligence (C4I) systems enable effective military awareness, decision-making and operation. By also integrating surveillance and reconnaissance (C4ISR) capabilities, intelligence analysis benefits of additional information regarding adversary assets and capabilities, in peacetime and conflict.
AI can enhance the decision-making process both at strategic, operational, and tactical operation level.

This Use Case proposes AI to support decision-making in a tactical operation scenario where multiple threats participate with different impacts and speeds, which require a quick response using effective countermeasures.

AI tools find a promising application to:

- data collection, correlation, and fusion from multiple platforms/sensors/probes in single or multi-domain environment (land, maritime, air, space, cyber).
- data extraction providing the information of interest (e.g. adversary assets), in particular when the amount of data collected is huge, for tactical operation applications and for intelligence analysis purposes.
- threat evaluation and weapon assignment, assisting the decision-making process to select the best defence resources against the threat in that timeframe.

Identified Impacts

- **Impact on the system:** These applications can have a positive impact on system performance (TAID-02:O, TAID-21:O, PER-03:O) and trustworthiness of AI (ASDC-01:O). AI technology has the potential to improve future C4I/C4ISR providing faster decisions, reducing the system response time (PER-01:O). However, the resources required for designing and developing such system and also for updating could be onerous (DCLC-02:R, DCLC-05:R). In addition, the reliability of the system needs to be ensured to prevent failure during mission (TAID-22:R).
- **Impact on the mission:** Military operations/specifics are impacted by improvement of goals achievement, mission situation assessment, reaction time and mission efficiency (OODA loop) (MOP-02:O, MOP-04:O, MOP-07:O). A better army management and allies' coordination can lead to a gain in supremacy vs. opponents at field (MSP-04:O, MSP-05:O). The missions can be improved by better assessment of own and opponents' intents, nonetheless they can be also compromised in case of system breach (MSP-06:N MSP-07:N).
- **Impact on the operator:** The fact that AI can handle a huge amount of data coming from a variety of sensors, platforms, systems operating in different domains, decreases human operator effort (MOP-06:O). However, challenges arise in the management and control of AI capabilities (TAID-11:R, HF-06:R) when ensuring that they are effective and up to date according to the changing needs typical of a distributed, evolving and contested environment (HUV-02:R, DCLC-02:R, DCLC-05:R). Relevant aspects also impacting the operator's activities are the need for having a clear and human-interpretable view of the system status provided by the AI models, so the decisions to be taken are well understood by the operator and the entailed consequences bearable (HF-08:R, TAID-12:R).

Involved Stakeholders

According to the approach presented in Section 2.4.1, this Use Case is on a **tactical level**. Therefore, the expected scenario takes place on a **short** time scale. Accordingly, the following main stakeholders are involved in this Use Case:

- AI customers
  - Legal responsibility
  - Technical responsibility
  - AI operators
    - Operator
    - Operator team members
    - Operators interacting with AI system
    - C4I operators
    - Commander training team
- AI partners
  - AI trainers

  - AI subject matter experts
- AI providers
  - Software platform providers
  - AI product providers
- AI producers
  - AI developers
- AI authorities
  - Executive power (Chain of command)

8.4. Impact Analysis Outcome

Based on the analysis of the Use Cases and on the expert's knowledge, the following subsections include a selection of the most relevant impact characteristics to consider when addressing trustworthiness for AI in Defence. For further detail, see "TAID Annex 04 AI Use Cases for Defence".

8.4.1. Impact of AI on the Sovereignty of Defence Systems

The concept of sovereignty is key in defence and the introduction of new technologies like AI should be evaluated through this perspective. The first European initiative to study the impact on sovereignty of the use of AI technology to develop military systems and support engineering processes has been done within the context of EICACS project [33]. The high-level sovereignty principles that applied in the past and still apply today are:

- Sovereignty relates to the ability of EU and its Member States to exercise their autonomy or self-determination.
- Sovereignty presumes the ability of EU and its Member States to independently analyse, decide and act.
- Organizations and individuals subject to EU and its Member States' jurisdiction are entitled to self-determination.
- Competent Jurisdictions define boundaries for EU and its Member States to exercise its sovereignty.
- Sovereignty shall be based on fundamental values, rights and principles and European and National regulations.

8.4.1.1 Definition of Sovereignty

In the literature, it is possible to find several general definitions of the sovereignty. In the CEN/CENELEC Workshop Agreement 17995 [71], a refinement of such general definitions have been proposed to address Digital Sovereignty and Technological sovereignty.

- **Sovereignty** is the ability of a country to autonomously analyse (understand/assess a situation), decide and act accordingly (those lead to the notions of autonomy of assessment, autonomy of decision, autonomy of action with a transverse notion of autonomy of governance).
- **Digital Sovereignty** is the ability to perform or support a function based on digital resources which include but are not limited to, data, information, software, processes, digital knowledge, human resources, hardware, digital infrastructure, engineering methods and tools.
- **Technological sovereignty** and Digital Sovereignty, while strongly overlapping (on hardware, infrastructure, engineering methods and tools) also differ in that, for example,

technological sovereignty includes non-digital technologies.

Based on the sovereignty principles above, and the definitions from the state of the art, it is possible to define the ***Sovereignty*** *in the context of* ***AI-based system development and usage as:***
Sovereignty in the relations between States signifies independence regarding a portion of the globe, along with the exclusive right to exercise, to the exclusion of any other State, the functions of a State. This encompasses self-reliance in development, secure and resilient supply chains, robust cybersecurity, and adherence to legal and ethical frameworks. It implies the safeguarding of sensitive data, the protection of critical digital and physical infrastructures, and the vigilance against external dependencies and cyber threats.

8.4.1.2 Main Characteristics of Sovereignty

The development of military systems developed with Artificial Intelligence (AI) is governed by a multifaceted set of sovereignty characteristics established in EICACS project [33]:

Table 7 - List of Characteristics defining Sovereignty

ID	Sovereignty Characteristic	Description
SC-01	Technological Independence	Self-sufficiency in developing and maintaining military AI to avoid dependency on external entities and to safeguard European and/or National security interests.
SC-02	Security and Digital Integrity	Protect AI systems and digital infrastructure against cyber threats.
SC-03	Legal and Ethical Frameworks	The use of military AI should comply with European, National and/or international laws, including adherence to ethical norms, human rights, and humanitarian rules.
SC-04	International Compliance and Dependencies Awareness	While engaging in international agreements and to avoid vulnerabilities, EU and its Member States should minimize external dependencies and have an appropriate level of control on remaining necessary ones, particularly with regard to foreign investments and technologies.
SC-05	Import and Export Control	EU and its Member States should regulate the trade of military AI technologies to prevent undesirable proliferation and to maintain their security autonomy.
SC-06	Indispensability and Dispensability	EU and its Member States should promote and support AI capabilities that are indispensable to allies and maintain a dispensability approach to sourcing to avoid single points of failure.
SC-07	Openness and Interoperability	Military AI systems should be compatible with various systems and technologies, enabling a dynamic and adaptable defence ecosystem. The development of common standards and sharing best practices with allies is regarded as beneficial.
SC-08	Infrastructure Sovereignty	Control over the essential AI development and deployment infrastructure as well as control over data generated and used is necessary to maintain sovereignty in military capabilities.
SC-09	Economic and Workforce Considerations	Development should be economically viable, and EU and its Member States should invest in education and training for a skilled workforce capable of creating and managing AI systems.

8.4.1.3 Risk-based approach to evaluate Sovereignty

To manage the potential impact of AI technologies on the sovereignty of defence systems, EICACS project [33] has proposed to adopt a risk-based approach to manage sovereignty requirements. A risk-based approach is suitable in this context because it would be too restrictive to simply forbid the usage of a given technology not satisfying all sovereignty characteristics (listed above) when it is possible to identify technical or procedural mitigation means to reduce the sovereignty risk under a certain acceptable threshold.

For example, this risk-based framework may provide sufficient argument to authorize the use of a non-European AI technology with regards to sovereignty risks given that an efficient and independent monitoring function is able to detect any unintended behaviour at AI constituent level or at system level (e.g., detection of a change in the expected behaviour of the AI -base system to do an adversarial attack enabled by a vulnerability of the AI technology used).

A first list of Sovereignty Hazards (SH) has been defined within the frame of EICACS project [33] to support this risk-based approach.

Table 8 - Sovereignty Characteristics and Associated Hazards

ID	Sovereignty Characteristic/TAID Property	Sovereignty Hazard (SH)
TAID-06	Autonomy	SH-1: Risk of AI System acting outside of its assigned intended decision-making autonomy scope, leading to unintended engagements or escalation of conflict that are not no longer controllable by human oversight or operational rules adaptation.
SC-01	Technological Independence	SH-2: Dependence on foreign technology could lead to supply chain vulnerabilities and potential exploitation by adversaries.
SC-02	Security and Digital Integrity	SH-3: Cyber threats could compromise AI systems, leading to data breaches, loss of sensitive information, or manipulation of AI actions.
SC-03	Legal and Ethical Frameworks	SH-4: AI deployment in conflict with legal and ethical standards could lead to war crimes, civil liability, and international condemnation.
SC-04	International Compliance	SH-5: Non-compliance with international agreements may result in sanctions or diplomatic disputes, undermining national interests.
SC-05	Import and Export Control	SH-6: Unauthorized proliferation of military AI technology could lead to an arms race or empower adversaries.
TAID-03	Resilience	SH-7: Inadequate system resilience could result in loss of military capability in the face of disruptions, cyberattacks, or other failures.
SC-06	Indispensability	SH-8: Overemphasis on being indispensable could lead to overreliance by allies and a lack of support in times of need.
SC-06	Dispensability	SH-9: A lack of diversified sources could lead to strategic vulnerability if a single critical supplier fails.
TAID-04	Protection	SH-10: Failure to protect critical systems and values may result in undermining national security and democratic processes.
SC-07	Openness and Interoperability	SH-11: Lack of interoperability could isolate the national military systems and prevent effective coalition operations.
SC-08	Infrastructure Sovereignty	SH-12: Loss of control over AI infrastructure could lead to an inability to maintain or operate military AI systems autonomously.
SC-09	Economic Considerations	SH-13: Mismanagement of AI system development costs could lead to budget overruns and unsustainable economic burdens.
SC-09	Workforce Development	SH-14: Insufficiently trained personnel may lead to operational failures or accidents in the deployment and use of military AI systems.
TAID-01	Accountability	SH-15: Lack of clear accountability could result in misuse of AI, erosion of public trust, and difficulty in addressing failures.

8.4.2. Impact of AI on the Trustworthiness Assurance Strategy

In this section, the impacts of AI on the trustworthiness assurance strategy for military systems are discussed. The term assurance should be read as the planned and systematic actions necessary to provide adequate confidence and evidence that a military system satisfies its intended function.

8.4.2.1 The dynamic nature of Operational Design Domain in the Defence Sector

The Operational Design Domain (ODD) defines the specific operating conditions under which an AI constituent within a given system is intended to function. For civilian use cases, such as commercial aircraft, the ODD is generally expected to remain static due to the highly regulated nature of these environments. Civil aviation operates within strict parameters, with flight paths, behaviours, and reactions to specific situations clearly defined by international and national regulations. Changes to the CONOPS (Concept of Operations) in civil aviation are typically infrequent and result from considerable planning and regulatory approval. The predictability of the operating environment allows for a static ODD, where parameters like weather conditions, air traffic, and airport operations are within known limits and changes are methodically introduced and communicated well in advance.

In military use cases, however, the ODD "as operated" can be much more dynamic due to the unpredictable and adversarial nature of the environment. Military operations often involve strategic manoeuvres and tactics such as concealment and deception to gain advantage over

an adversary. An enemy might employ concealment manoeuvres designed specifically to mislead and manipulate the performance of adversary military systems. Consequently, military systems must be able to adapt their ODD to respond to such tactics and maintain operational effectiveness. This may involve adjusting to new patterns of enemy behaviour, adapting to changes in the environment that are intentionally induced by the enemy, or responding to the presence of novel threats.

As illustrated in Figure 13 below, the expectation for ODD adaptation in military systems is driven by the need for flexibility and resilience in the face of an intelligent and adaptive adversary. This stands in contrast to civilian systems that operate in a more predictable and stable environment, where changes to the ODD are the result of a deliberate and controlled process.

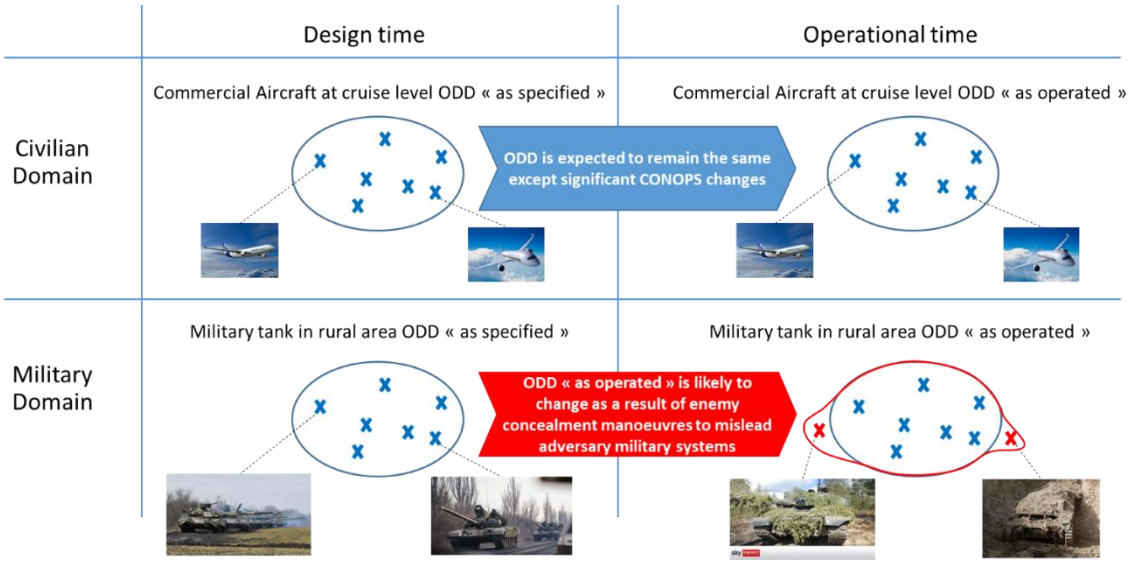


Figure 13 - Differences between civilian and military ODDs

8.4.2.2 Evolution of the balance between development assurance and runtime assurance

Traditionally in system engineering (including software and hardware), the term *assurance* defines the planned and systematic actions necessary to provide confidence and evidence that a system or a product satisfies given requirements. A process is needed which establishes levels of confidence that development errors that can cause or contribute to identified failure conditions (feared events defined by a safety/security/human factor assessment) have been minimized with an appropriate level of rigor. This henceforth is referred to as the *development assurance* process [13].

When the system is deployed in service, *runtime assurance* refers to a set of techniques and mechanisms designed to ensure that a system behaves correctly during its execution [72]. This involves monitoring the system's behaviour in real-time and taking predefined actions to correct or mitigate any deviations from its expected performance, safety, or security requirements. Runtime assurance can be particularly important in critical and/or autonomous military systems where failures could lead to significant harm or loss.

There are several key components and strategies often associated with runtime assurance [73]:

- **Runtime Monitoring:** Observing the system's operational behaviour and performance to detect anomalies, errors, or deviations from its specified/expected behaviour and domain (ODD).
- **Assertion Checking:** Using assertions or conditions that the system must satisfy at

specific points during execution. If an assertion fails, it indicates a potential error or issue.

- **Recovery Strategies:** Predefined procedures or actions that the system automatically takes when certain types of errors or anomalies are detected. These can range from simple error logging and alerts to more complex recovery mechanisms like switching to a backup system or reducing functionality to maintain safety.
- **Adaptive Behaviour:** In some cases, runtime assurance mechanisms can adjust the system's behaviour in response to changing operational conditions or detected issues to maintain performance and safety requirements.
- **Safety Mechanisms:** Incorporating features designed to prevent unsafe conditions or mitigate their effects should they occur. This can include watchdog timers, fail-safes, and redundancy.

Runtime assurance is a dynamic and ongoing process that contrasts with design-time verification methods, which aim to ensure the correctness of a system before it is deployed. While design-time techniques are essential for building reliable systems, runtime assurance provides an additional layer of protection by addressing issues that may not have been foreseen during the development phase or that arise from the system's interaction with the real world.

In Figure 14, the evolution of the balance between development assurance and runtime assurance is illustrated. One can observe that the introduction of AI technologies and autonomy capabilities tips the balance towards greater needs of runtime assurance to compensate the increasing difficulty to perform comprehensive a priori development assurance activities.

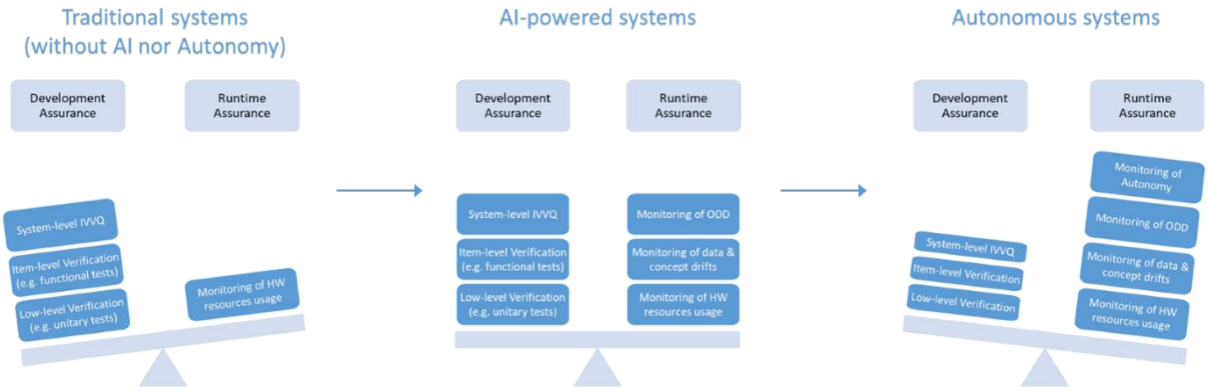


Figure 14 - Evolution of the balance between development assurance and runtime assurance

8.4.3. Impact of data frugality on the trustworthiness for AI in the Defence Domain

When referring to the Defence domain, the availability of relevant data poses an additional challenge, as security matters are also under thoughtful consideration to avoid any undesirable disclosure of sensitive information. European initiatives have emerged to support research activities on this specific field [74]. In the following sub-sections, the main attributes of the impact of the frugality of data on the trustworthiness for AI in the defence domain are discussed.

8.4.3.1 Importance of data on AI Trustworthiness

AI models are highly dependent on a vast amount of accurate, complete, timeliness, and representative data in order to train efficiently any AI-architecture apart from the lab performance validation. Implementing robust Data and AI Governance practices is essential to ensure data quality, integrity, and compliance throughout the AI development process. AI systems based on Machine Learning (ML) and Deep Learning (DL) techniques, usually rely on massive quantities of annotated training data to achieve high performance, which is also closely linked to the well-defined representation of the learning context. Data and AI Governance frameworks help establish clear policies for data collection, annotation, and usage, mitigating risks associated with bias and ensuring lawful, ethical, and reliable considerations are addressed.

Most of the state-of-the-art ML techniques provide high accuracy and impressive results under well-defined scopes and clearly annotated datasets. However, this is not the case when there is a lack of labelled representative data or missing classes of interest, which further deteriorate the outcomes of such approaches, resulting in poor performances, under conditions that could be considered normal for a human but not included in the training dataset. Effective Data Governance can help identify and address these data gaps, ensuring more comprehensive and representative datasets.

Besides the need for collecting and labelling the required data, data preparation is equally critical as it refines and transforms the collected data into a usable format. It improves model convergence and accuracy, balances the datasets to avoid bias, and eliminates redundant information. Data Governance practices play a crucial role in standardizing data preparation processes, ensuring consistency across different AI projects, and maintaining data lineage for transparency and accountability. Effective data preparation guided by strong Data Governance practices, ensures consistency, saves time and resources by preventing issues during model training, and leads to more accurate and robust ML models.

8.4.3.2 Nature of data on the Defence Domain

Military data are often considered as scarce, incomplete, or too specific. Collecting large scale training datasets in many military applications could be infeasible, as for certain fields only very limited samples can be acquired which usually are not disclosed introducing many restrictions. Another aspect of the specificity of Defence Domain on the available datasets, is relying on the fact that the nature of the missions requires a fast adaptation on new environments, making existing training datasets obsolete. To overcome such constraints and maximize the insertion and eventually the impact of AI in the defence domain, AI techniques that incorporate basic frugal principles may act as potential solutions.

8.4.4. Model Deception

AI deception methods (like painting tanks in pink colours) will never stop because human ingenuity will always come with some new deception idea, that the AI trainer did not think about. But this is not an argument for today, where fundamental AI capacities are still lacking, e.g. “moving from A to B with an autonomous land vehicle in a GPS denied natural environment”. For such elementary tasks, painting tanks in pink is a very remote concern. Focusing on AI enabled fundamental capacities, (that finally nobody at this stage may try to compromise), would be a reasonable target.
Development of such basic AI capabilities can be done only through massive testing on “open source” common platforms (both Hardware and Software).

8.4.5. Civilian & Defence Regulation: Dual Compliance of Military Systems

- Used in peace time: key driver

- EU pay a lot of attention there – Lisbon treaty
- How to enable the use AI in systems when in peace time?
- How to train operators to the military use?
- Consider the uses that are only military

8.4.6. SW vs AI systems / Evolution of development practices with focus on real-time/embedded solutions

The integration of the AI within the defence sector is inevitably leading to a paradigm shift in the way SW is developed and executed, with distinct impacts on embedded, real-time, safety-critical systems and offline-executed SW.
In the realm of embedded SW, it is necessary to account for more stringent requirements than other applications. For example, a Design Assurance Level (DAL) is assigned to each capability deployed in airborne systems, to assure the highest level of safety and reliability. In addition, HW must be robust enough to withstand all operating conditions, instead for offline-executed SW, HW does not need to withstand such a wide range of operating conditions.
In case of AI applications, system development begins with inputs and desired results and the goal is to create a solution that emulates the underlying physical behaviour with a high degree of confidence. In general, the advent of AI changes how SW requirements, design, and verification processes are managed.

AI application demands the identification of the suitable neural network, known as Neural Architecture Search (NAS). Hence, SW design for AI cannot merely rely on functional decomposition aimed at achieving optimal modularization and function encapsulation.
For both embedded and offline-executed SW, when a trained model and its inference must be installed on a specific HW, and the inference does not meet the HW constraints, additional techniques, such as pruning and quantization, must be applied to reduce resource needs. This approach should be regulated to prevent inference deviations from the trained model, which could adversely impact the verification phase.

AI solutions require peculiar validation techniques, especially in the learning phase, as explained in Figure 8 from the EASA Concept of Design Assurance for Neural Network (CODANN II [75]). Once the inference is available, the verification phase can apply standard SW methodologies.

Given the high computational complexity and memory requirements of AI solutions, it is expected that equipment for embedded solutions will evolve to include Multi-Core Processor (MCP) technologies first, but also Graphics Processing Units (GPUs) in the future. In both cases, homologation and certification present a challenge, particularly in demonstrating determinism and preventing crashes in the equipment.

Despite these challenges, improvements in AI can be achieved using well-established criteria and a critical approach. Standard methods and AI solutions can still be adopted together to model functions, ensuring a robust and reliable SW solution for both embedded and offline-executed systems.

8.5. Discussion and Perspectives

8.5.1. Way-forward to manage AI Impacts on Sovereignty

To improve the trustworthiness of AI-powered European Defence Systems, the risk-based approach presented above need to be further refined, using lower-level metrics and indicators to measure on a given AI technology (e.g., learning framework such as Scikit-Learn, Tensorflow, Pytorch, etc.), the level of sovereignty and associated risks and develop an automated pipeline to support this evaluation.

8.5.2. Way-forward to manage AI Impacts on the Trustworthiness Assurance Strategy

The following research topics will contribute to progress in terms of knowledge and practical mean to manage AI Impacts on the Trustworthiness Assurance Strategy:

- **Distributed and Decentralized Runtime Assurance:** For systems operating in a distributed or decentralized environment, such as swarms of drones or autonomous vehicles in traffic, runtime assurance must be effective across the entire network. This area explores how to coordinate assurance mechanisms across multiple entities and how to ensure collective safety and performance.
- **Runtime Assurance for Trusted Dynamic Delegation of Authority in Autonomous Systems:** This research topic focuses on the challenge of ensuring trusted mechanisms for the delegation of authority from human operators to autonomous systems, including the dynamic and context-aware transfer of control and decision-making capabilities. It addresses the core issue of how and when an autonomous system should be allowed to make decisions, take actions, or extend its operational design domain (ODD) without direct human intervention.
- **Human-in-the-Loop Assurance:** Investigating how human operators can be effectively integrated into the runtime assurance process, especially for systems where human oversight is crucial. This includes developing interfaces and protocols for human-machine collaboration to manage complex or ambiguous situations that automated systems might not handle well on their own.
- **Security and Runtime Assurance:** Researching ways to integrate runtime assurance with cybersecurity measures to protect against attacks that could compromise the safety and reliability of AI-enabled and autonomous systems. This involves detecting and mitigating security threats in real-time, including those involving adversarial AI techniques.
- **Resource-Aware Runtime Assurance:** Developing runtime assurance mechanisms that are sensitive to the resource constraints of the system, such as computational power, memory, and energy. This is crucial for ensuring that assurance processes do not overly burden the system's ability to perform its primary functions.

8.5.3. Way-forward to manage data frugality

Novel and innovative models and architectures start to emerge to overcome the obstacles raised by the scarcity and lack of appropriate training data in the Defenced domain, and usually are categorized based on two core pillars:

1. Data augmentation techniques that produce synthetic data to account for the needed datasets. Synthetic data comprise information that is produced in an algorithmically manner, rather than generated in real events. The simulated data can be exploited in the training process of the AI models; however, quantifiable indicators of their performance need to be identified to ensure the accuracy and representativeness of real field data.

2. Development of AI models that do not require significant amount of data to be trained. Frugal models reuse existing models or transfer learning and continuous learning principles, which can contribute to improved performances with less data, reuse models trained under different domains e.g. civil protection and continuously training the model when external data are incorporated to the models, respectively.

8.5.4. Way-forward to manage AI integration in embedded systems

Considering that in aeronautics the need for platform certification has a long history, a good starting point could be the objectives defined in the EASA Concept Paper: “First usable guidance for Level 1 machine learning applications” [29]. These objectives need to be refined and integrated with the processes outlined in the RTCA/DO-178C [36] adopted by FAA, EASA and Transport Canada for software development. A clear and comprehensive definition of these objectives will not only provide a robust framework for the development and for the acceptance of AI solutions by certification authorities.

Furthermore, the definition of software standards is likewise crucial to ensure that software artifacts meet the set objectives. These standards could be based on AI W-shaped lifecycle (Section 5.2.3) and consider further enhancements, particularly in the area of data management. Undisciplined data management, which neglects essential steps such as data cleaning, integration, feature selection, transformation, and reduction, can compromise system reliability and, consequently, trust in the system.

Moreover, embedded systems add another layer of complexity due to their main focus on efficient resource management, for instance to operate deterministically and provide uninterrupted services, like in case of car’s drive-by-wire or ABS (anti-lock braking system) that need to be fail-proof. This demand might come at the expense of explainability and portability that are necessary for trustworthiness. To mitigate this issue, SW industry is developing software processes, standards, and tools enhancing interoperability and hardware optimization. For instance, ONNX (Open Neural Network Exchange) provides a common set of operators - the building blocks of machine learning and deep learning models - and a common file format.

Lastly, hardware is indeed the final challenge to consider. AI often requires significant computational power and a large amount of RAM. Determining the Worst-Case Execution Time (WCET) in MCPs and GPUs is particularly challenging due to potential interferences among cores sharing resources. In this context, it is crucial for industries and organizations to establish together solid best practices and methodology. Techniques such as Bandwidth Allocation and Monitoring (BAM), among others, are going to be explored, investigated, and refined to guarantee the achievement of this requirement.

9. Conclusions & Recommendations

9.1 Wrap-up

This whitepaper has been elaborated by the TAID Working Group to support EU Members States and Defence Industry to better prepare, plan and develop the future AI systems. To achieve this goal, multidisciplinary expert knowledge was gathered in order to analyse the different dimensions and aspects involved in the adoption and integration of emergent AI technology in the Defence sector. In the aim to provide a comprehensive view, the following subjects were covered:

- AI definition and Taxonomy
- Identification of stakeholders for AI in defence
- Testing & Evaluation, Validation and Verification standards, tools and methodologies
- Human Factors
- Ethical AI considerations
- Impact analysis of AI in Defence
- Military Use cases and scenarios for Trustworthiness of AI

- Recommendations and next steps

First, the taxonomy and stakeholders' topics provided essential terminology and structure to understand and position the process of integrating AI technology within the different Stakeholders of AI in Defence. The legal analyses provide foundation for alignment with policies, regulations and standards in scope. An extensive list of standards addressing different technical concerns (like AI safety) was also included, given the key role of standardisation during systems design and development. Yet being a complex subject, the evaluation of trustworthiness of systems integrating AI technology was stated in terms of a list of so named Trustworthiness AI properties for Defence.

To facilitate the validation and verification of TAID properties, some guidance was provided, including methods and tools to operationalize the process by following a suggested end-to-end engineering life cycle. Additionally, the ethical and human factors analyses produced a comprehensive list of ethical recommendations and human factors requirements to ensure ethics and human wellbeing.

To assess the impact of integrating AI technology, a multi-characteristics method was introduced covering different dimensions, ranging from human-organizational to technical aspects. The method was illustrated by evaluating impact (risks, opportunities, neutral) of a set of ongoing Use Cases, paying special attention to Defence specifics. Finally, as a result of the impact analyses and based upon expert knowledge, a prioritized list including most relevant impact characteristics was elaborated and discussed. The paper is concluded by incorporating a series of recommendations in the aim to better prepare the next steps towards trustworthy integration of AI in Defence systems.

9.2 Recommendations to EU Defence organizations and MS

AI systems will play an increasingly significant role in future military applications. As AI systems differ from existing rule-based algorithms and systems, it is important that AI systems are also trustworthy for future users. A key aspect of creating trust is the systematic validation of AI systems under relevant operating conditions. In the authors' view, this requires product-neutral evaluations of AI systems regarding their potential as well as their weaknesses, based on standards relevant to military systems.

9.2.1 Managing AI Impacts on European/Member States Sovereignty

To manage the impacts of AI on sovereignty, it is recommended to develop specific metrics and indicators that can assess the sovereignty risks of various AI technologies, particularly learning frameworks like Scikit-Learn, TensorFlow, and PyTorch, for the development of AI-enabled mission critical military systems. These metrics and indicators should help quantify sovereignty and associated risks using a systematic evaluation framework shared by all European defence actors and a common database of sovereign AI technologies maintained by EDA. Additionally, the implementation of a Data and AI Governance Framework is crucial to continuously evaluate, monitor and assess risks and their impacts, ensuring that AI systems remain trustworthy. Finally, the current risk-based approach should be refined to include a detailed analysis of AI components, with a focus on mitigating sovereignty risks at every stage of development.

9.2.2 Establishing a European AI Risk Repository for Defence

It is recommended to build an European AI Risk Repository, inspired by the EU AI Act [1], ALTAI guidelines [35], Information Security frameworks, MIT AI Risk Repository [76], and the NIST AI Risk Repository, to systematically document and assess risks associated with AI

technologies, ensuring alignment with Data and AI Governance. This repository would serve as a central knowledge base, categorizing AI risks by industry, use case, and potential impacts and mitigations. It would provide policymakers, developers, and stakeholders with practical guidance on managing these risks, offering strategies to mitigate vulnerabilities specific to AI deployment in defence and other critical sectors. Additionally, the repository should be continuously updated, through an appropriate change management process, with real-world data and insights from ongoing projects, ensuring it remains a relevant and actionable resource for European AI innovation and governance.

9.2.3 Managing the Transition to more Runtime Assurance for AI-enables Systems

To enhance the trustworthiness of AI systems, the following research areas need to be prioritized.

First, it is essential to improve runtime assurance mechanisms for AI-enabled systems operating in distributed environments, such as drone swarms, to ensure collective safety and consistent performance across all entities.

Second, developing reliable methods for dynamic delegation of authority from human operators to autonomous systems is critical, especially in situations where AI-enabled systems need to make decisions without direct human intervention.

Third, integrating human oversight into the runtime assurance process is necessary to handle complex situations where AI-enabled systems may not perform optimally. Lastly, runtime assurance should incorporate cybersecurity measures to safeguard AI-enabled systems against potential adversarial threats, while remaining mindful of system resource constraints like computational power and energy.

9.2.4 Data Governance and Data Frugality

In response to data limitations in the defence sector, innovative solutions should focus on two core areas. On top of the on-going analysis of the defence data space concept that aims to provide safe and secure access to defence data another two aspects should be considered. First, synthetic data generation techniques need to be adopted to create additional datasets for AI training. However, it is crucial to develop quantifiable performance indicators to ensure that these synthetic datasets are realistic and representative of real-world conditions. Second, the development of frugal AI models that require minimal data for training is essential, or models that can be reliably trained on synthetic data. Techniques such as transfer learning and continuous learning should be leveraged to improve model performance, allowing models trained in one domain (e.g., civil protection) to be adapted for defence with ongoing adjustments as new data becomes available.

To address these data challenges in the defence sector, a comprehensive Data and AI Governance Framework is essential. This framework integrates people, processes and technology and establishes the policies, procedures, standards, regulations, and tools, necessary for effectively managing an organization’s data assets. Furthermore, by actively engaging stakeholders at all levels, this framework promotes a culture of accountability, continuous improvement, and responsible AI use. As a result, it ensures AI systems are developed and deployed in accordance with Trustworthy AI.

This framework encompasses several key components. Firstly, the definition of roles and responsibilities ensures accountability within the organization. Secondly, data modelling,

aiding in understanding relationships among different domains, entities, and attributes. Moreover, this structured approach not only enhances understanding but also supports interoperability, allowing different systems to communicate and share data seamlessly.

Additionally, data lineage and traceability are also critical, as they track origins and transformations of data, ensuring data quality and integrity. In addition, successful metadata management enhances usability by contextualizing data. For instance, this includes creating Data dictionaries and glossaries to standardize terminology, crucial for ensuring interoperability among NATO Allies. Furthermore, implementing data catalogs helps manage data assets and facilitate discovery.

Moreover, these resources will help classify privacy and security requirements for attributes, enabling effective access and sharing of information. All together, these elements play a vital role in supporting data and AI governance, risk management, and data security and privacy.

In conclusion, a strong Data and AI governance strategy is essential for aligning with business objectives, driving value and fostering a culture of data stewardship across the organization. Ultimately, this approach enables strategic, tactical and operational efficiency, facilitating better-informed decision-making, and enhancing trust in data, all while ensuring compliance with regulations.

9.2.5 Enabling AI Integration in Embedded Systems

To ensure successful AI integration in embedded systems, it is possible to benefit from the work performed by EASA Concept Paper Issue 2 [29] with the necessary changes/enhancements to operate a military aircraft and weapon systems. These military **standards** should ensure AI-based solutions are acceptable to certification authorities. Moreover, enhancing data management practices through robust Data and AI Governance is vital for maintaining system reliability and trustworthiness. This includes ensuring thorough specific **processes** for data integration, storage, cleaning, and transformation. Lastly, a collaborative approach is needed to optimize hardware resources for embedded systems. AI technologies often require significant computational power, so establishing robust **methodologies** for managing hardware constraints is critical for ensuring system performance without sacrificing explainability or trust (e.g. Energy capacity).

9.2.6 Incremental Change Management Implementation for AI-Enabled Systems

To ensure the safe and effective deployment of AI models in mission-critical environments, it is recommended to establish a Data and AI Governance Framework for the incremental qualification of AI-based systems. Incremental qualification refers to the need for defence use cases to progressively update AI models (e.g. to handle enemy manoeuvres) at a much faster rate than for traditional civilian use cases (e.g. a model updated overnight to prepare for the next mission the next day).

This approach allows for incremental updating and deployment of new AI models, assessing risks and impacts while ensuring they meet operational constraints, such as reliability, performance, and real-time responsiveness. The framework should define clear stages for testing and validating AI models in controlled environments before gradual integration into live missions, thus maintaining operational continuity and safety. This process would also incorporate feedback loops from real-world missions to improve model accuracy and adaptability over time, ensuring that updated AI systems remain mission-ready and trustworthy.

9.2.7 Develop an AI Risk Management Framework for Defence

An essential step to validate an AI system is defining the risks throughout the whole life cycle. In the “TAID Annex 01 Risks Analysis”, there is a list of sample risks regarding data, security, application, ... that can be used as a starting point for an evaluation of the AI system. For each risk there is a correlated possible damage, some suggested mitigation methods and related impacted properties. A sample list of AI system properties with metrics for measuring each of them is presented in the “TAID Annex 02 Trustworthiness Properties”. To help assessing the trustworthiness of the system there are different toolkits and frameworks available on the market, and in the “TAID Annex 03 Toolkits and frameworks”, some of these are detailed and can be choose or used as a reference of what’s available on the market.

9.2.8 Develop an End-to-End Standardized Evaluation Framework for Generative-purpose AI used in Defence Applications

It is recommended to develop an end-to-end standardized evaluation framework for General-purpose AI used in defence applications, such as mission planning, threat analysis, and decision support.

This end-to end framework should assess Large generative AI models using a risk-based approach and by considering the definition of potential mitigations means at system level including the protection functions like monitoring and specific procedures based on human oversight. This end-to-end standardized evaluation framework will also be based on specific criteria including accuracy, transparency, robustness, and compliance with ethical and legal requirements.

For instance, in mission planning, large generative AI models like Large Language Models (LLMs) must be evaluated for their ability to provide reliable and context-sensitive recommendations without bias or misinformation (where models generate incorrect or misleading information often called hallucinations), ensuring explainability (clear understanding of how decisions are made), and maintaining data security (protecting sensitive military information). Addressing these challenges is crucial to prevent misinterpretation of data or actions based on faulty recommendations, which could lead to mission failure or security risks.

This risk-based end-to-end evaluation framework for defence applications may benefit from existing frameworks such as the EU AI Act framework for high-risk AI systems which should comply with standards for transparency, accountability, and human oversight. For example, large generative AI models used in autonomous decision-making for defence should be subject to strict risk management and data governance protocols that may be inspired from those mandated by the EU AI Act but adapted to defence context and objectives, ensuring they are both trustworthy and compliant with applicable defence regulations and standards.

9.2.9 Development of Use Cases integrating AI-technology for Defence

Increasing AI-technology readiness is a process tightly coupled to the development of Use Cases. Whereas the Use Cases analysis provided in sections 8.1 to 8.3 provides initial guidance, it mainly illustrates salient elements, like AI technological traits, stakeholders and balance between risks and opportunities. To have a more comprehensive view and perspective on the specific AI technology maturity, their usages, and the entailed risks and opportunities, it is recommended to foster projects targeting the development of Use Cases integrating AI technology in Defence. Referred projects should help to increase Use Case maturity, determine viability/deploy-ability, and demonstrate alignment amongst the characteristics and requirements in scope.

9.2.10 Consolidation of multidisciplinary teams to ensure effective alignment between human-value and AI-technology characteristics

Analysing the impact of integrating AI-technology in Defence systems is crucial to identify opportunities and risks. Whereas Sections 8.1 **Error! Reference source not found.**and 8.2 p resent a method to conduct a preliminary impact assessment and a detailed list of characteristics, multiple factors can still influence the process and produce undesirable biases. Ensure effective and balanced impact analysis of projects integrating AI technology is a complex subject that calls for strengthening and fostering the development of multidisciplinary teaming including but not limited to ethicists, engineers, regulators, authorities, etc. It also demands the development of approaches, including methods and frameworks, to ensure effective alignment between human-value (high-level) characteristics and technical (low-level) characteristics.

9.2.11 Human Factors & Ethics Recommendations

As highlighted in Chapters 6 and 0 the following summary recommendations are:

An ALTAI for Defence:

Given the criticality of ethics to both Defence Operations and the use AI, it is imperative that there is a common means of reference and guidelines for all operations throughout the design lifecycle. ALTAI [35] and FCAS have made commendable in-roads, however, a defence-specific version in which both systems and value-based engineering approaches is required. Ethics must be applied in practice and cannot remain as a set of abstract concepts, lists or tick-box exercises that stakeholders do not fully understand. All stakeholders across the design lifecycle must understand how their role and actions as an individual and team member impacts the preservation of life and peace for current and future generations. Likewise, the Defence forces at domain, national and/or international level must also understand how the ethical values, decision-making and actions required to uphold international humanitarian law and support all the humans in the wider socio-technical system. On top of the what to add when how (where applicable) for the ALTAI-based questions.

Defence specific ethical guidance:

It must be contextualized so that all stakeholders understand how ethical practice is relevant to their job role and operations- they must understand how to use the guidelines in practice and when to use them. This can be achieved using a scenario-based approach, use-cases, or even serious games. They key factors here are 1) context, timing, and application of the guidelines. 1) Context: Guidelines should be developed by multi-disciplinary teams, led by an expert or value lead, and should be presented in a format that relevant stakeholders can understand and apply the ethical values in their work practices, under normal and non-normal conditions. 2) Timing: The guidelines should be referenced and applied at every stage in the design lifecycle. 3) Ethical practice does not only apply to those in the field during combat. They also apply to those responsible for procurement, technology design, training, evaluation, certification, and review. All stakeholders should be trained to understand the ethical values and how to apply them specific to their domain and role. Presenting them with a list, a set of questions or application to use (such as ALTAI [35]) without training by a relevant expert is not sufficient for defence operations.

Value-Based Engineering:

Develop and define “Ethical Values for Defence”, and process to be integrated in the whole AI Lifecycle Engineering. It is essential to develop and define “Ethical Values for Defence”, and for this process to be integrated in the whole AI Lifecycle Engineering. These guidelines should

guide stakeholders in identifying what other values are and should be for their own operational domain. This process must be led by an accredited expert - a “value lead” - to ensure that a full understanding is reached by stakeholders and that the guidelines are applied to the full lifecycle. The values can and should complement requirements and other existing values. The “value lead” can support stakeholders in linking the ethical values as part of current/future metrics and KPIs to ensure overlap with evaluation and review processes.

Testing and Evaluation methodology/framework for the Ethical requirements:

Many of the current means testing and evaluation do not account for the introduction of new technologies, procedures, training etc. This is also true for ethical requirements. Consideration must be given as to how ethical requirements can be tested and evaluated in practice. This must be carried out by a multi-disciplinary team and lead by an accredited expert /Value lead. If ethics are indeed given their due importance by defence forces and the international community, there must be a means of measuring if and how well they have been deployed withing the system. Collaboration of experts in ethics, value-based engineering and human factors should be encouraged to assess how such a framework can be derived, what the metrics/KPIs should be and how they should be audited on national and international levels.

Trust in the System:

All humans operating must know that they will be supported by their team and respective organization. If operators are to fully engage with new technologies and non-human team members, it is critical that there is a fundamental level of trust. From a systems perspective, human operators should be able to operate under the umbrella of a just culture especially if it is demanded of them to make decisions in complex, fast-changing, safety-critical environments. This does not apply to malpractice. It is necessary that defence forces at national and international level provide a duty of care to those carrying out operations on their behalf, so that they know they will be treated fairly and supported well by their peers and superiors. It is in the interests of all (current and future generations) that operators make the best decisions in a timely manner- this is far-less likely to occur if they suspect that they will be treated unjustly for decisions made or do not trust in the system, technologies, or team members around them. It would be prudent to research how **just culture improvements** have benefitted the wider system in other safety critical environments with a view to how this could be achieved for defence operations using AI.

9.2.12 Standardization Management Plan for AI standards for Defence

The identification of Trustworthy AI standards in the present whitepaper was the initial effort to introduce the related AI standards already applicable or under development in the civil sector. A broader Standardization Management Plan and EDSTAR (European Defence Standards Reference System) are EDA supportive functions that can further assess and propose standards applicable for AI in defence. This effort performed mainly by a dedicated Expert Group could come up with selection of “Best Practice” Standards for Defence and to identify potential gaps or needs for new standards development.

9.2.13 AI Taxonomy for Defence Update

The already publicly available OSRA Defence Technology Taxonomy [3] includes a dedicated AI Section reflecting a first attempt to use a common terminology for defence. However, in the context of the present work it has been identified that there are some missing elements especially for trustworthiness of AI and therefore it is recommended to take action on its update process.

9.2.14 Testing and Evaluation Infrastructure Requirements for AI-based defence systems

In order to further facilitate the Testing and Evaluation actions for AI-based systems in defence there is a need to assess and develop a list of suitable test centres that will be able to support and perform the advanced needs of AI testing. EDA with the Defence Test and Evaluation Base offers a supportive function with an already established network of more than 300 military test centres across EU that may assist the future T&E actions of AI-based systems. Tailored frameworks and methodologies along with specific infrastructure requirements for testing and evaluation of AI-based systems are topics proposed for next step action lines in order to provide to MS and Defence Industry available solutions compliant with the selected standards for AI systems in defence.

Appendixes

Appendix 1 - Common AI Taxonomy Terms

- AUTOMATION: Pertaining to a process or system that, under specified conditions, functions without human intervention.
- CONNECTIONIST MODEL: Form of cognitive modelling that uses a network of interconnected units which generally are simple computational units.
- DEEP LEARNING: AI approach to creating rich hierarchical representations through the training of neural networks with many hidden layers.
- GENERAL PURPOSE AI MODEL: it is an AI model trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market (acc. EU AI Act [70]).
- GENERAL-PURPOSE AI SYSTEM: it is an AI system which is based on a general-purpose AI model, and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems (acc. EU AI Act [70]).
- INFERENCE: reasoning by which conclusions are derived from known premises. In AI, a premise is either a fact, a rule, a model, a feature, or raw data (The term "inference" refers both to the process and its result).
- INPUT DATA: Data provided to or directly acquired by an AI system based on which the system produces an output.
- LANGUAGE MODEL: A language model is an approximative description that captures patterns and regularities present in natural language and is used for making assumptions on previously unseen language fragments.
- LARGE GENERATIVE AI MODEL: a typical example for a general-purpose AI model, given that it allows for flexible generation of content, such as in the form of text, audio, images, or video, that can readily accommodate a wide range of distinctive tasks (acc. EU AI Act [70]).
- MACHINE LEARNING (ML): Machine Learning is a branch of artificial intelligence (AI) and computer science which focuses on development of systems that are able to learn and adapt without following explicit instructions imitating the way that humans learn, gradually improving its accuracy, by using algorithms and statistical models to analyse and draw inferences from patterns in data.
- MACHINE LEARNING ALGORITHM: Algorithm to establish parameters, according to a given criteria of a machine learning model from data.
- ML MODEL: A parametrized function that takes features as input and predicts labels or make decisions as output. The parameters are defined in the training process. Typical phases of an AI model's workflow are: Data collection and preparation, Model development, Model training, Model accuracy evaluation, (Hyper)parameters' tuning, Model usage, Model maintenance, Model versioning.
- MODEL TRAINING: Process to establish or to improve the parameters of a machine learning model, based on a machine learning algorithm, by using training data.
- MODEL VERIFICATION: Confirmation through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled.
- NATURAL LANGUAGE: Language which is or was in active use in a community of people, and the rules of which are mainly deduced from the usage. Natural language is any human

language, which can be expressed in text, speech, sign language etc. such as English, Spanish, Arabic, Chinese, or Japanese[8].

- NATURAL LANGUAGE GENERATION: Task of converting data carrying semantics into natural language.
- NATURAL LANGUAGE PROCESSING (NLP): The ability of a machine to process, analyse, and mimic human language, either spoken or written. Discipline concerned with the way computers process natural language data.
- NATURAL LANGUAGE UNDERSTANDING: Extraction of information, by a functional unit, from text or speech communicated to it in a natural language, and the production of a description for both the given text or speech, and what it represents.
- NEURAL NETWORK: A computer system inspired by living brains, also known as artificial neural network, neural net, or deep neural net. It consists of two or more layers of neurons connected by weighted links with adjustable weights, which takes input data and produces an output. Whereas some neural networks are intended to simulate the functioning of biological neurons in the nervous system, most neural networks are used in artificial intelligence as realizations of the connectionist model.
- NEURON: AI primitive processing element which takes one or more input values and produces an output value by combining the input values and applying an activation function on the result. Examples of nonlinear activation functions are a threshold function, a sigmoid function, and a polynomial function.
- PARAMETER: ML internal variable of a model that affects how it computes its outputs. Examples of parameters include the weights in a neural network, or the transition probabilities in a Markov model.
- PREDICTION: output of a machine learning model when provided with input data.
- REINFORCEMENT LEARNING: A type of ML in which the algorithm learns by interacting with an environment and acting toward an abstract goal, such as “earn a high video game score” or “manage a factory efficiently.” During training, each effort is evaluated based on its contribution toward the goal.
- TASK: Actions required to achieve a specific goal. These actions can be physical or cognitive (examples of tasks include classification, regression, ranking, clustering, and dimensionality reduction).
- TRAINING DATA: Subset of input data samples used to train a machine learning model.

[8] To be distinguished from programming and formal languages, such as Java, Fortran, C++, or First-Order Logic.

Appendix 2 - Discussion on Safety-critical vs. Mission-critical

Safety is a broad concept that applies to different domains, each with its own specific definitions and contexts. In the perimeter of AI for defence applications, a suitable general definition of “System Safety” could be: “the application of engineering and management principles to achieve functional safety of electrical, electronic, and programmable electronic systems. It involves identifying, assessing, and mitigating risks to ensure that the system performs its intended functions safely and reliably, even in the presence of potential faults or failures." (borrowed from IEC 61508 [10]).

In Civil Aviation domain, safety is “the state in which risks associated with aviation activities, related to, or in direct support of the operation of aircraft, are reduced and controlled to an acceptable level” (ICAO definition [77]). According to the standard SAE ARP-4754 [13], safety is strongly coupled to airworthiness, which refers to the certification process of an aircraft or system indicating that it is safe to fly and meets the applicable regulatory requirements. Safety plays a crucial role in achieving airworthiness by enabling regulatory compliance, mitigating risks, and ensuring safe operation.

In Military Aviation domain, the most suitable definition of safety is the one provided in the standard MIL-STD-882E [78]: “Freedom from conditions that can cause death, injury, occupational illness, damage to or loss of equipment or property, or damage to the environment.” Along with this general concept, the more concrete definition of “System Safety” is as follows: “The application of engineering and management principles, criteria, and techniques to achieve acceptable risk within the constraints of operational effectiveness and suitability, time, and cost throughout all phases of the system life-cycle.”.

The commonalities of System Safety definitions across different standards in different domains emphasize a consistent approach that includes the use of systematic engineering practices, lifecycle risk management, compliance with safety regulations, and the final objective of ensuring that systems operate safely under all foreseeable conditions.

Safety-critical according to MIL-STD-882E [78] is “A term applied to a condition, event, operation, process, or item whose mishap severity consequence is either Catastrophic or Critical (e.g., safety-critical function, safety-critical path, and safety-critical component).” This definition highlights the importance of identifying and managing items within a system that, if they fail, could lead to significant safety hazards.

According to MIL-STD-882E [78] Mission-critical is a term which applies to items whose failure could result in the inability to accomplish a mission, or the loss of a critical capability.

Thus, the terms “safety-critical” and “mission-critical” can be confused, however they refer to different topics. In order to illustrate this difference, the mission-critical characteristic of “aircraft survivability” can be used.

According to [79], “Aircraft Survivability is the capability of an aircraft to avoid or withstand a manmade hostile environment.” Aircraft survivability refers to the capability of an aircraft to withstand and operate effectively in hostile environments (i.e. non-peacetime condition), particularly when subjected to enemy threats such as anti-aircraft weapons, missiles, electronic warfare (EW), and other forms of attack. Survivability encompasses various design features and characteristics such as low detectability (e.g. stealth technology), damage resistance (e.g. bullet-resistant armouring plating), countermeasures (e.g. chaff & flares or laser jamming of infrared seekers), or redundancy (e.g. different sources of geolocation to mitigate the risk of GPS signal jamming).

On the other hand, safety typically involves the introduction of mitigation means to fulfil with the applicable safety requirements at different levels (i.e. Aircraft, System, or Item), such as safe/arm switches to avoid unintended activation of military features on-ground (e.g. missile firing) or other design provisions to deal with general hazards during peacetime operation.

The main difference between both terms lies in the nature of the risks considered. Around safety, hazards can be considered accidental or fortuitous, whereas, in survivability, hazards are intentional and characteristic of a hostile environment.

Mission-critical in military aviation refers to capabilities that are essential for the planning and execution of tactical missions and contributing to its success. Taking this into account the aircraft survivability can be considered as a mission-critical capability, and thus not contributing to the overall safety of the aircraft.

Appendix 3 - Addition to the Toolkits topic

Ensure conformity to the AI Act [70] is not yet a straightforward task. Since the AI Act is implemented through the New Legislative Framework (NLF) it relies on standards to give the conformity means to the industry. Then it is up to the mandated body (in this case the CEN CENELEC[9]) to produce the required harmonised standards. Any means of conformity has to be related to the requirement that the CEN CENELEC is about to published in 2025. Therefore, some of the current conformity assessment frameworks and tools are speculating on the content of such standards. That said, some standards ought to be central in the work of the CEN CENELEC. For example, the ISO/IEC 42001[10], ISO/IEC 42005[11], ISO/IEC 29119-11[12], ISO/IEC 25059[13], ISO/IEC 24029-2[14] are probably going to be used as an acceptable framework to build the harmonized standards.

Some tools are starting to implement the requirements from such standards. The OECD has started to list existing tools that matches these standards, however this list is not yet taking into account their relevancy for the harmonized framework of CEN CENELEC since it is not yet stable. The list can be found here online[15].

[9] https://www.cencenelec.eu/
[10] https://www.iso.org/standard/81230.html
[11] https://www.iso.org/standard/44545.html
[12] https://www.iso.org/standard/79016.html
[13] https://www.iso.org/standard/80655.html
[14] https://www.iso.org/standard/79804.html
[15] https://www.oecd.org/en/publications/tools-for-trustworthy-ai_008232ec-en.html

Appendix 4 – TAID Working Group Members

The Trustworthiness for AI in Defence Working Group consists of multidisciplinary volunteers from different entities (eg. MODs, Industry, Academia, RTOs) and with different expertise covering both technical and ethical/legal aspects.

For the whitepaper development the initial material used derived from the EDA's "Trusted AI and Standardization Workshop" that took place on Sep 2023 and collected input from industry and academia (50 proposals) in response to the workshop's subject where more than 200 participants attended the event.

The main workshop's outcome was the establishment of the TAID working group to support and contribute to EDA's action item to develop a whitepaper/report that would reflect the workshop's outcome and the analysis performed on the selected topics in order to form the scope and the relevant actions for Trustworthiness for AI in Defence.

The whitepaper development included three phases: whitepaper drafting with 6 sub-groups working in parallel, harmonisation phase and editing phase.

The list with the Governmental and Institutional representatives of the working group is:

Full Name	Entity
Isidoros Monogioudis	EDA
Sebastiano Maruca	EEAS
Marcilli Gianluca	IT MOD
Francisco Lamas Lopez	ES MOD
Daniele Bet	IT MOD
Luciana Morogan	RO MOD
Luis Javier Costa Giraldo	INDRA
Rebeca Blasco Jiménez	INDRA
Dr. Andreas de Jonge	DE JONGE GmbH
Bruno Carron	AIRBUS DE
Alison M. Kay	TCD IE
Janaina Ribas De Amaral	AIRBUS DE
Gabriel Pedroza	ANSYS
Fateh Kaakai	THALES
Michel Barreteau	THALES
Dr. Andreas Tollkühn	MBDA DE
Liisa Janssens	TNO NL
Yvonne Hofstetter	21 STRATEGIES
Chrystèle Johnson	MDBA FR
Monika Venckauskaite	BPTI
Alessio Cavallin	LEONARDO

Full Name	Entity
Fabio Magosso	LEONARDO
Javier Ferrero Micó	GMV
Dr. Markus Hosbach	IABG
Antonio Monzón-Diaz	AIRBUS DE
Joseph Machrouh	THALES
Dr. Frank Beer	INFODAS
Alexis De Cacqueray	AIRBUS DE
Andromachi Papagianni	CERTH
Christophe Guettier	SAFRAN ELECTRONICS & DEFENSE
Papantoniou Vassilios	HTR
Ruben Post	TNO NL
Tuulia Timonen	CGI
Kintzios Spyridon	CERTH
Natalia Giogiou	CERTH
Jari Turkia	CGI
Adrien Becue	THALES
Patricia Besson	THALES
Jonas Stiensmeier	RHEINMETTAL
Mauro Patini	LEONARDO
Roman Sztyler	ESG
Simona Soare	Lancaster University

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
AIS	Autonomous and Intelligent Systems
AI RMF	AI Risk Management Framework
ALTAI	Assessment List for Trustworthy Artificial Intelligence
AQAP	Allied Quality Assurance Publications
ARP	Aerospace Recommended Practice
ART	Adversarial Robustness Toolbox
ASIL	Automotive Safety Integrity Levels
AWI	(ISO/IEC) Approved Work Item
C2	Command and Control
C4I	Command, Control, Communications, Computers, and Intelligence
C4ISR	Command, Control, Communications, Computers Intelligence, Surveillance, and Reconnaissance
CD	(ISO/IEC) Committee Draft
CEN	European Committee for Standardization
CENELEC	European Committee for Electrotechnical Standardization
CI/CD	Continuous Integration and Continuous Delivery
CLC	Short for CENELEC
CODANN II	(EASA) Concept of Design Assurance for Neural Network II
CONOPS	Concept Of Operations
DAL	Design Assurance Level
DCLC	Development Cycle/Life Cycle
DCS	Data-Centric Security
DEF STAN	UK Defence Standards
DevOps	(software) Development and (IT) Operations
DIS	ISO/IEC Draft International Standard
DL	Deep Learning
DTS	(ISO/IEC) Draft Technical Specification
EASA	European Union Aviation Safety Agency
ED	EUROCAE Document
EDA	European Defence Agency
EDF	European Defence Fund
EICACS	European Initiative for Collaborative Air Combat Standardisation
EMACC	European Military Airworthiness Certification Criteria
EMAR 21	(EDA) European Military Airworthiness Requirements (2021 edition)
EO	Electro-Optical (sensors)
EN	(UNE) European Norm
EUROCAE	European Organization for Civil Aviation Equipment

Abbreviation	Definition
ETSI	European Telecommunications Standards Institute
EVR	Ethical Value Requirements
EW	Electronic Warfare
FAA	US Federal Aviation Administration
FCAS	Future Combat Air System
FDIS	(ISO/IEC) Final Draft International Standard
FRA	EU Agency for Fundamental Rights
GDPR	General Data Protection Regulation
GPS	Global Positioning System
GPU	Graphics Processing Unit
GR	(ETSI) Group Report
GRC	Governance, Risk, and Compliance
HF	Human Factors
HLEG	High-Level Expert Group
HMI	Human Machine Interface
HMT	Human-Machine Teams
HPC	High Performance Computing
HUV	Human-World Values
HW	Hardware
IC2E	International Conference on Cloud Engineering
ICAO	International Civil Aviation Organization
IEA	International Ergonomics Association
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IEHF	Institute of Ergonomics and Human Factors
IHL	International humanitarian law
ILO	International Labour Organization
ISO	International Organization for Standardization
IT	Information technology
IVVQ	Independent Verification, Validation, and Qualification
JTC	Joint Technical Committee
KPI	Key Performance Indicator
MBSE	Model-Based Systems Engineering
MCP	Multi-Core Processor
MHC	Meaningful Human Control
MIL-SPEC	Military Specifications
MIL-STD	Military Standards
ML	Machine Learning
MLOps	Machine Learning Operations
MOD	Ministry of Defence
MOP	Military Operation
MS	Member State
MSP	Military Specifics

Abbreviation	Definition
NAS	Neural Architecture Search
NASA	National Aeronautics and Space Administration
NATO	North Atlantic Treaty Organization
NB	Nota Bene
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NMAA	National Military Airworthiness Authority
OD	Operational Domain
ODD	Operational Design Domain
OECD	Organisation for Economic Co-operation and Development
ONNX	Open Neural Network Exchange
OODA	Observe, Orient, Decide and Act
OSRA	Overarching Strategic Research Agenda
PAS	(ISO/IEC) Publicly Available Specification
PER	System Performance
PRU	(NATO) Principles of Responsible Use
PUF	Physically Unclonable Function
PWI	(ISO/IEC) Preliminary Work Item
FaRADAI	Frugal and Robust AI for Defence Advanced Intelligence
RL	Reinforcement Learning
RTO	Research and Technology Organization
SAE	Society of Automotive Engineers
SC	Sovereignty Characteristic
SH	Sovereignty Hazard
SIL	Safety Integrity Levels
SQuaRE	Systems and software Quality Requirements and Evaluation
STANAG	(NATO) Standardization Agreement
STS	Socio-Technical Systems
SW	Software
TAI	Trustworthiness of AI
TAID	Trustworthiness for AI in Defence
TARA	Threat Assessment and Remediation Analysis
TR	(ISO/IEC/ETSI) Technical Report
TS	(ISO/IEC) Technical Specification
TRL	Technology Readiness Level
UAV	Unmanned Aerial Vehicle
UC	Use Case
UN	United Nations
UNE	Una Norma Española
VUCA	Volatile, Uncertain, Complex, Ambiguous
VBE	Value-Based Engineering
XAI	Explainable Artificial Intelligence

Bibliography

[1] European Commission, “Artificial Intelligence Act (Regulation (EU) 2024/1689), Official Journal version of 12 July 2024”.

[2] OECD, “Explanatory Memorandum on the updated OECD definition of an AI System,” 2024.

[3] EDA, “OSRA Defence Technology Taxonomy v2.0. [Online]. Available: https://eda.europa.eu/docs/default-source/documents/osra-defence-technology-taxonomy.pdf,” 2021.

[4] ISO/IEC, “22989:2022 Information Technology Artificial Intelligence Artificial Intelligence Concepts and Terminology,” ISO, 2022.

[5] ISO/IEC, “23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML),” 2022.

[6] ISO/IEC, “TR 24029-(1-2-3):2021 Artificial Intelligence (AI); Assessment of the robustness of neural networks,” 2021.

[7] IEEE Standard Association, “IEEE 7010-2020 - IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being,” New York, 2020.

[8] IEEE Standard Association, “IEEE P7012 Standard for Machine Readable Personal Privacy Terms”.

[9] IEEE Standard Association, “7001-2021 IEEE Standard for Transparency of Autonomous Systems,” 2021.

[10] IEC, “61508:2010 Functional Safety Standard,” IEC, 2010.

[11] EN, “50129:2019 Railway applications - Communication, signalling and processing systems - Safety related electronic systems for signalling,” 2019.

[12] ISO, “26262:2018 Road vehicles; Functional safety,” 2018.

[13] SAE International, “ARP-4754B - Guidelines for Development of Civil Aircraft and Systems,” SAE, 2023.

[14] ISO/IEC, “DIS 12792 Information technology; Artificial intelligence; Transparency taxonomy of AI systems,” (under development).

[15] ISO/IEC, “24028:2020 Information technology; Artificial intelligence; Overview of trustworthiness in artificial intelligence,” 2020.

[16] IEEE Standard Association, “7000:2021 Standard Model Process for Addressing Ethical Concerns during System Design,” 2021.

[17] ISO/IEC, “42001:2023 Information technology; Artificial intelligence; Management system,” 2023.

[18] L. Janssens, “Clarifying Military Advantages and Risks of AI Applications via a Scenario,” in *The Quest for AI Sovereignty, Transparancy and Accountability - Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition*, Kyoto, Japan, United Nations Internet Governance Forum, 2023, pp. 197-213.

[19] Treaty on European Union, “Consolidated version of the Treaty on European Union OJ C 202, 7.6., p. 17,” 2016.

[20] OECD, “Recommendation of the Council on Artificial Intelligence,” 2024.

[21] NATO, “Summary of NATO's revised Artificial Intelligence (AI) strategy. [Online]. Available: https://www.nato.int/cps/en/natohq/official_texts_227237.htm,” 2024.

[22] NATO, “Summary of the NATO Artificial Intelligence Strategy. [Online]. Available: https://www.nato.int/cps/en/natohq/official_texts_187617.htm,” 2021.

[23] European Commission - High Level Expert Group on Artificial Intelligence, “Ethics Guidelines for Trustworthy AI,” 08 04 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai. [Accessed 15 10 2024].

[24] M. M. M. B. T. &. V. M. Mäntymäki, “Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance,” (arXiv:2206.00335)., 2022.

[25] NIST, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” 2023.

[26] ISO/IEC, “23894:2023 Information technology; Artificial intelligence; Guidance on risk management,” 2023.

[27] NIST, “Trustworthy and Reponsible AI - Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,” 2024.

[28] EASA, “Artificial Intelligence Roadmap 2.0 - A human-centric approach to AI in aviation,” 2023.

[29] EASA, “EASA Artificial Intelligence concept paper (proposed Issue 2) open for consultation,” [Online]. Available: https://www.easa.europa.eu/en/newsroom-and-events/news/easa-artificial-intelligence-concept-paper-proposed-issue-2-open. [Accessed 10 10 2024].

[30] SAE International, “AIR6988: Artificial Intelligence in Aeronautical Systems: Statement of Concerns,” 2021.

[31] EUROCAE/SAE, “ED-324/ARP-6983 - Process Standard for Development and Certification/Approval of Aeronautical Safety-Related Products Implementing AI,” SAE International, (under development).

[32] Airbus; Fraunhofer FKIE, “White paper - The Responsible Use of Artificial Intelligence in FCAS – An Initial Assessment,” 2024. [Online]. Available: https://www.fcas-

forum.eu/en/articles/responsible-use-of-artificial-intelligence-in-fcas/. [Accessed 15 10 2024].

[33] EU EICACS Project, “EU EICACS website,” EU EICACS Project, [Online]. Available: https://www.eicacs.eu/. [Accessed 10 10 2024].

[34] EU AI4DEF Project, “EU AI4DEF Website,” EU AI4DEF Project, [Online]. Available: https://ai4def.com/. [Accessed 10 10 2024].

[35] European Commission - High-Level Expert Group on Artificial Intelligence, “The Assessment List for Trustworthy Artificial Intelligence,” 17 July 2020. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment. [Accessed 15 10 2024].

[36] EUROCAE/RTCA, “DO-178C - Software Considerations in Airborne Systems and Equipment Certification,” RTCA Incorporated, 2011.

[37] EUROCAE/RTCA, “DO-254 - Design Assurance Guidance for Airborne Electronic Hardware,” RTCA Incorporated, 2000.

[38] IEA, "What is Ergonomics/ (Human Factors/Ergonomics)?," 2024.

[39] ICAO, "Manual on Human Performance (HP) for Regulators, Do 10151, First Edition," 2021.

[40] J. Schraagen, "Responsible Use of AI in Military Systems (1st ed.). Chapman and Hall/CRC.," 2024.

[41] Peeters et al, "Hybrid collective intelligence in a human–AI society. AI and Society, 36(1), 217–238," 2020.

[42] CIEHF, Chartered Institute of Ergonomics & Human Factors, White Paper: "Learning from adverse events," 2020.

[43] Woods, David, "Limits of Automata—Then and Now: Challenges of Architecture, Brittleness, and Scale." Journal of Cognitive Engineering and Decision Making, 2024.

[44] Malcolm MacLachlan, (ed.): "Maritime psychology: research in organizational & health behavior at sea," Springer, pg.8, (ISBN 9783319454283), 2017.

[45] McVeigh, J. et al, "Psychosocial and organisational aspects of work at sea and their implications for health and performance," In: Textbook of maritime health (3rd Ed.). Bergen : Norwegian Centre for Maritime and Diving Medicine, 2022.

[46] IEA/ILO, "Principles and Guidelines for HF/E Design and Management of Work Systems." Joint Document by IEA and the International Labour Organization (ILO)., 2021.

[47] Salmon, P.M. and Read, G.J.M., "Using principles from the past to solve the problems of the future: Human factors and sociotechnical systems thinking in the design of future work." Hum Factors Man, 28: 277-280, 2018.

[48] Salmon et al, "Pilot error versus sociotechnical systems failure: a distributed situation awareness analysis of Air France 447," Theoretical Issues in Ergonomics Science, vol. 17, no. 1, pp. 64-79, 2016.

[49] Woods. D, "Resolving the Command-Adapt Paradox: guided Adaptability to Cope with Complexity." In book: Compliance and Initiative in the Production of Safety: A Systems Perspective on Managing Tensions & Building Complementarity. Chapter: 8, Springer Briefs in Safety Management, 2018.

[50] Kay A. and McDonald, N., "Future Flight Operations: Communication, Collaboration and Resource Management," Transport Transitions: Advancing Sustainable and Inclusive Mobility - Proceedings of the 10th TRA Conference, Dublin, Ireland. Springer, 2024.

[51] Roberts et al, "State of Science: models and methods for understanding and enhancing teams and teamwork in complex sociotechnical systems." Ergonomics 65(2): 161-187, 2022.

[52] Laux, J., "Institutionalised Distrust and Human Oversight of Artificial Intelligence: Toward a Democratic Design of AI Governance under the European Union AI Act," 2023.

[53] European Commission, "Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," Pub. L. No. COM, 2021.

[54] Mcdonald, N., Kay, A., Morrison, R. and Ryan, M., "How Automation May Transform the Ways in Which Crew Managet Peak Workload and Incapacitation." 19th International Symposium on Aviation Pscyhology. Dayton, Ohio, U.S.A., 2017.

[55] Mcdonald, N., Kay, A., Morrison, R., Ryan, M. and Zon, R., "A Validated Description of How Crew Manage Flight Operations for Two-Pilot and Reduced Crew Operations." The First International Symposium on Human Mental Workload. M.C. a. L. Leva, L. Dublin, 2017.

[56] Mcdonald, N., Kay, A., Liston, P., Morrison, R. and Ryan, M., "An Integrated Framework for Crew-Centric Flight Operations." Human Computer Interaction International, Los Angeles, U.S.A, Springer, 2015.

[57] Stanton et al, "Distributed situation awareness in dynamic systems: theoretical development and application of an ergonomics methodology." Ergonomics. 10-22;49(12-13):1288-311, 2006.

[58] Kay, A. et al, "Case study in RAF Boeing E3D Sentry. Modelling Command and Control- event Analysis of Systematic Teamwork." N. A. Stanton, Baber, C. and Harris, D. Hampshire, U.K., Ashgate: 157 - 180, 2008.

[59] Stewart et al, "Distributed Situational Awareness in Airborne Warning and Control Aircraft: application of a novel ergonomics methodology." Cognition, Technology and Work 10: 221 - 229., 2008.

[60] Harris et al, "Report of the Working Group to Identify Future Challenges Faced by the Implementation of Resource Management in Remote and Distributed Teams," Lecture Notes in Computer Science(), vol 14692. p. 190-200 Springer, Cham., 2024.

[61] Ward et al, "A Case Study of a Whole System Approach to Improvement in an Acute Hospital Setting." Int. J. Environ. Res. Public Health, 19, 1246, 2022.

[62] Geary et al, "A socio-technical systems analysis of the application of RFID-enabled technology to the transport of precious laboratory samples in a large acute teaching hospital," Applied Ergonomics, Volume 102, 2022.

[63] McDonald, N. et al, "A Mindful Governance model for ultra-safe organisations," Safety Science, 120, p753 - 763, 2019.

[64] Harris, D., "Single-pilot airline operations: Designing the aircraft may be the easy part." The Aeronautical Journal: 1-21., 2023.

[65] FRA, "Bias in Algorithms, Artificial Intelligence and Discrimination," European Union Agency for Fundamental Rights, Vienna, 2022.

[66] FRA, "Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights," Vienna, 2019.

[67] S. Spiekermann and T. Winkler, "Value-based Engineering for Ethics by Design," *Elsevier SSRN,* pp. 1-23, 12 May 2020.

[68] ISO/IEC, "ISO/IEC/IEEE 24748-7000:2022 - Systems and software engineering — Life cycle management," ISO/IEC, 2022.

[69] Eggert, L., "Rethinking ' Meaningful Human Control' in Schraagen, J.M., Responsible Use of AI in Military Systems (1st ed.)." Chapman and Hall/CRC, 2024.

[70] Council of the EU, "Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights," Brussels, 2022.

[71] CEN/CENELEC, "CEN/CENELEC Workshop Agreement - CWA 17995," 06 2023. [Online]. Available: https://www.cencenelec.eu/media/CEN-CENELEC/CWAs/RI/cwa17995_2023.pdf. [Accessed 10 10 2024].

[72] NASA, "NASA/TM–20220015734 - Runtime Assurance of Aeronautical Products: Preliminary Recommendations," NASA, California, 2023.

[73] NASA, "NASA/CR–2020-220586 - Run Time Assurance as an Alternate Concept to Contemporary Development Assurance Processes," NASA, USA, 2020.

[74] EU FARADAI Project, "EU FARADAI Website," EU FARADAI Project, [Online]. Available: https://faradai.eu/. [Accessed 10 10 2024].

[75] EASA, "EASA CoDANN II," [Online]. Available: https://www.easa.europa.eu/en/document-library/general-publications/concepts-design-assurance-neural-networks-codann-ii. [Accessed 10 10 2024].

[76] MIT, "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intellenge. Available: https://cdn.prod.website-

files.com/669550d38372f33552d2516e/66bc918b580467717e194940_The%20AI%20Risk%20Repository_13_8_2024.pdf," 2024.

[77] ICAO, "Global Aviation Safety Plan," 2017-2019. [Online]. Available: https://www.icao.int/safety/Documents/Doc%2010004.2017-2019%20edition.alltext.en.pdf. [Accessed 15 10 2024].

[78] Department of Defense, "MIL-STD-882E System Safety," 2012. [Online]. Available: https://mail.system-safety.org/Documents/MIL-STD-882E.pdf. [Accessed 15 10 2024].

[79] R. E. Ball, "The fundamentals of aircraft combat survivability analysis and design," USA: American Institute of Aeronautics and Astronautics, 2003.

[80] ISO/IEC, "24030:2021 Information technology Artificial Intelligence (AI) Use Cases," 2021.

[81] ISO/IEC, "TR 24372:2021 Overview of computational approaches for AI systems," 2021.

[82] ISO/IEC, "TS 4213:2022 Assesment of ML classification peformance," 2022.

[83] ISO/IEC, "NP 23282 Artificial Intelligence; Evaluation methods for accurate natural language processing systems," (under development).

[84] ISO/IEC, "AWI TR 42106 Information Technology; Artificial Intelligence; Overview of differentiated benchmarking of AI system quality characteristics," (under development).

[85] ISO/IEC, "AWI TS 29119-11 Software and systems engineering; Software testing; Part 11: Testing of AI systems," (under development).

[86] ISO/IEC, "DIS 5392:2024 Information technology; Artificial Intelligence; Reference architecture of knowledge engineering," 2024.

[87] Cen/CLC/TR, "17894:2023 Artificial Intelligence Conformity Assessment," (under development).

[88] ISO/IEC, "CD 42006 Information technology; Artificial intelligence; Requirements for bodies providing audit and certification of artificial intelligence management systems," (under development).

[89] ISO/IEC, "AWI TR 42103 Information technology; Artificial intelligence; Overview of synthetic data in the context of AI systems," (under development).

[90] ISO/IEC, "PWI 42102 Information technology; Artificial intelligence; Taxonomy of AI system methods and capabilities," (under development).

[91] ISO, "31073:2022 Risk management vocabulary," 2022.

[92] ISO/IEC, "FDIS 5338:2023 Information technology; Artificial intelligence; AI system life cycle processes," 2023.

[93] ISO/IEC, "5339:2024 Information technology; Artificial intelligence; Guidance for AI applications," 2024.

[94] ISO/IEC, “38507:2022 Information technology; Governance of IT; Governance implications of the use of artificial intelligence by organizations,” 2022.

[95] ISO/IEC, “25058:2024 Systems and software engineering; Systems and software Quality Requirements and Evaluation (SQuaRE); Guidance for quality evaluation of artificial intelligence (AI) systems,” 2024.

[96] ISO/IEC, “AWI TS 17847 Information technology; Artificial intelligence; Verification and validation analysis of AI systems,” (under development).

[97] ISO/IEC, “NP TS 22443 Information technology; Artificial intelligence; Guidance on addressing societal concerns and ethical considerations,” (under development).

[98] ISO/IEC, “PWI 42105 Information technology; Artificial intelligence; Guidance for human oversight of AI systems,” (under development).

[99] ISO/IEC, “24368:2022 Information technology; Artificial intelligence; Overview of ethical and societal concerns,” 2022.

[100] ISO, “31000:2018 Risk management; Guidelines”.

[101] ISO/IEC, “TS 8200:2024 Information technology; Artificial intelligence; Controllability of automated artificial intelligence systems,” 2024.

[102] ISO/IEC, “25059:2023 Software engineering; Systems and software Quality Requirements and Evaluation (SQuaRE); Quality model for AI systems,” 2023.

[103] ISO/IEC, “CD TS 6254 Information technology; Artificial intelligence; Objectives and approaches for explainability and interpretability of ML models and AI systems,” (under development).

[104] ISO/IEC, “TS 12791 Information technology; Artificial intelligence; Treatment of unwanted bias in classification and regression machine learning tasks,” (under development).

[105] ISO/CD, “PAS 8800 Road Vehicles; Safety and artificial intelligence,” (under development).

[106] ISO/IEC, “TR 5469:2024 Artificial intelligence; Functional safety and AI systems,” 2024.

[107] ISO/IEC, “AWI TS 22440 Artificial intelligence; Functional safety and AI systems; Requirements”.

[108] ISO/IEC, “38500:2024 Information technology; Governance of IT for the organization,” 2024.

[109] ISO/IEC, “27000:2018 Information technology; Security techniques; Information security management systems; Overview and vocabulary,” 2018.

[110] ISO/IEC, “TR 24027:2021 Information technology; Artificial intelligence; Bias in AI systems and AI aided decision making,” 2021.

[111] IEC, “61800-7-1:2015 Adjustable speed electrical power drive systems - Part 7-1: Generic interface and use of profiles for power drive systems - Interface definition,” 2015.

[112] ISO/IEC, "TR 27550:2019 Information technology; Security techniques; Privacy engineering for system life cycle processes," 2019.

[113] ISO/IEC, "27042:2015 Information technology; Security techniques; Guidelines for the analysis and interpretation of digital evidence," 2015.

[114] ISO/IEC, "27043:2015 Information technology; Security techniques; Incident investigation principles and processes," 2015.

[115] E. Union, "Vertrag über die Europäische Union (konsolidierte Fassung)," Brüssel, 2012.

[116] D. Silver, T. Hubert and J. Schrittwieser, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Sience,* vol. 362, no. 6419, pp. 1140-1144, 2018.

[117] D. Silver, A. Huang and C. Maddison, "Mastering the game of Go with deep neural networks and tree search," *Science,* vol. 529, pp. 484-489, 2016.

[118] O. Vinyals, I. Babuschkin and W. Czarnecki, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature,* vol. 575, pp. 350-354, 2019.

[119] J. Perolat, B. De Vylder and D. Hennes, "Mastering the game of stratego with model-free multiagent reinforcement learning," *Science,* vol. 378, no. 6623, pp. 990-996, 2022.

[120] D. E. Denning and B. J. Strawser, "Naval Postgraduate School Publications," 2015. [Online]. Available: https://faculty.nps.edu/dedennin/publications/Active%20Cyber%20Defense%20-%20Cyber%20Analogies.pdf. [Accessed 10 10 2024].

[121] NATO, "Data Centric Security Implementation Plan 2.0," NATO, 2022.

[122] K. Wrona, "Towards Data-Centric Security for NATO Operations," *Springer - Communications in Computer and Information Science,* vol. 1790, no. Digital Transformation, Cyber Security and Resilience, pp. 75-92, 2023.

[123] IEEE Standard Association, "2894-2024 IEEE Approved Draft Guide for an Architectural Framework for Explainable Artificial Intelligence," 2024.

[124] IEEE Standard Association, "P2976 Standard for XAI - eXplainable Artificial Intelligence - for Achieving Clarity and Interoperability of AI Systems Design," (under development).

[125] IEEE Standard Association, "P7011 Standard for the Process of Identifying and Rating the Trustworthiness of News Sources".

[126] IEEE Standard Association, "P2863 Recommended Practice for Organizational Governance of Artificial Intelligence".