Trustworthiness for AI in Defence (TAID)

-

Trustworthiness Properties Annex

Submission date: 09/05/2025

Document version: 1.0

Authors: TAID Working Group (TAID WG)

Table of Contents

1. System engineering /software derived Properties

This is the list of properties directly derivable from general systems and software engineering. Mostly oriented to define systems and systems of systems general properties for software application. This means that these properties should be used for any application that involves software development.

HW Capacity to support AI complexity [TAID-05]

REFERENCES: None.

DEFINITION: The ability of the HW to support the execution of an AI algorithm/application in a safe and efficient way.

METRICS: None.

Sustainability [TAID-32]

REFERENCES: None.

DEFINITION: More efficient usage of the energy budget of the system itself or in the energy spent for system design, production, supply chain, etc.

METRICS: None.

Controllability [TAID-11]

REFERENCES: EN ISO/IEC 22989:2022

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

Explainability [TAID-12]

REFERENCES: EN ISO/IEC 22989:2022

DEFINITION: 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.

Metrics: Estimate the degree of explainability of a piece of information by measuring its relevance to answering a (pre-defined) set of archetypal questions.[1]

Maintainability [TAID-19]

References: EN ISO/IEC 22989:2022

Definition: 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.

Metrics: Refer to the maintainability index or alternative maintainability models. [2]

Reliability [TAID-22]

References: EN ISO/IEC 22989:2022

Definition: Property of consistent intended behaviour and results that enables to provide required prediction, recommendation, and decision consistently correctly during its operation stage.

Metrics: Evaluate system based on the following cases:

- Performance on data similar to data used for training ("in-distribution"),
- Performance on data mostly similar to data used during training, but with some variation ("near-distribution"),
- Ability to identify data that is significantly different from data used during training ("out-of-distribution"),
- Quality of estimating prediction confidence ("uncertainty calibration"), and
- Robustness to adversarially altered inputs ("adversarial robustness").[3]

Repeatability [TAID-23]

References: EN ISO/IEC 22989:2022

[1] https://arxiv.org/abs/2109.05327v5

[2] Heitlager I, Kuipers T, Visser J (2007) A practical model for measuring maintainability. In: 6th International conference on the quality of information and communications technology (QUATIC 2007) pp 30–39. IEEE

[3] Why Improving AI Reliability Metrics May Not Lead to Reliability | Center for Security and Emerging Technology (georgetown.edu)

DEFINITION: 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.

METRICS: Using standard deviation.[4]

Reproducibility [TAID-25]

REFERENCES: Harald Semmelrock, Simone Kopeinik, Dieter Theiler, Tony Ross-Hellauer, and Dominik Kowald. 2023. Reproducibility in Machine Learning-Driven Research. arXiv preprint arXiv:2307.10320 (2023).

DEFINITION: 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.

METRICS: Using a scientific method and quantifying degrees of reproducibility.[5]

Using standard deviation.[6]

Robustness [TAID-28]

REFERENCES: EN ISO/IEC 22989:2022

DEFINITION: Ability to maintain their level of performance, as intended by their developers, under any circumstances.

METRICS: Measure delays, errors, deviations, introduce errors and simulate attacks.[7]

Testability [TAID-33]

---

[4] Douglas A. Milikien. “Measuring Reproducibility and Repeatability of an AI-based Quantitative Clinical Decision Support Tool Having a Medical Decision Point.” PharmaSUG 2022 – Paper MD-174.

[5] Gundersen, Odd Erik and Sigbjørn Kjensmo. “State of the Art: Reproducibility in Artificial Intelligence.” *AAAI Conference on Artificial Intelligence* (2018).

[6] Douglas A. Milikien. “Measuring Reproducibility and Repeatability of an AI-based Quantitative Clinical Decision Support Tool Having a Medical Decision Point.” PharmaSUG 2022 – Paper MD-174.

[7] https://arxiv.org/abs/2203.12048v1

References: EN IEC 62628

Definition: 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.

Metrics: Testers can and should be able to control the input data fed into the AI system to observe how it responds under various conditions. This includes providing both typical and atypical inputs, modifying input features, introducing noise or perturbations, and testing different input scales or ranges.

Testers can and should be able to control the configuration of the AI model itself. This includes adjusting hyperparameters, the overall model architecture, or feature selection to assess the system's sensitivity to variations.

Testers can and should be able to control simulations or emulations of specific conditions or environments that are relevant to the AI system or more specifically, to the algorithms underlying that system.

Testers can and should be able to generate synthetic or simulated data to control the characteristics, distribution, or complexity of various types of inputs. This enables targeted testing of specific scenarios that may be difficult to encounter or reproduce in real-world data, often for a variety of reasons that say nothing about how common or not those scenarios are in a real-world context.[8]

Transparency [TAID-36]

References: EN ISO/IEC 22989

Definition: 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 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.

Metrics: When calculating an AI transparency score, it is important to consider several key factors. These factors provide critical insights into the decision-making processes and underlying mechanisms of AI systems. They include:

[8] The Spectrum of AI Testing: Testability – Stories from a Software Tester (testerstories.com)

1. Documentation: The availability of comprehensive documentation detailing the AI system's architecture, algorithms, and data sources is crucial for transparency. Organizations should provide clear and accessible documentation to users and stakeholders.

2. Algorithmic Explanations: AI systems should offer explanations for their decisions or recommendations. Techniques such as interpretability methods, natural language generation, or rule-based systems can be employed to provide interpretable explanations.

3. Data Sources and Preprocessing: The transparency of the data used to train AI models is essential. Organizations should disclose the sources, quality, and potential biases present in the data. Additionally, documenting the preprocessing steps taken to prepare the data for AI training is essential.

4. Model Interpretability: AI models should be interpretable to understand how inputs are transformed into outputs. Techniques like model-agnostic interpretation or rule-based models can enhance interpretability.

5. Bias and Fairness Analysis: Assessing and mitigating biases in AI systems is critical to ensure fairness. Evaluating the presence of biases and actively working to eliminate them demonstrates a commitment to transparency and ethical AI.[9]

Resilience [TAID-03]

REFERENCES: EN ISO/IEC 22989

DEFINITION: Resilience is the ability of the system to recover operational condition quickly following an incident.

METRICS: Resilience can be specified by measuring the MTTR (Mean time to recovery) of the system.

Model Correctness [TAID-41]

REFERENCES: ED-324/ARP6983 (draft 5b)

DEFINITION: Ability of a model to maintain its level of performance under all nominal (not processed by the model robustness) conditions within the ML ODD (Operational Design Domain).

METRICS: Measure the model Accuracy.

[9] How to Calculate AI Transparency Score | Blog (playerzero.ai)

Dependability [TAID-42]

References: ISO/IEC/IEEE 15026-1:2019

Definition: Ability to perform as and when required.

Metrics: Measure failure rates and number of faults. Refer to EN 62628:2012-09.

Governability [TAID-15]

References: NATO PRU (Principles of Responsible Use)

Definition: 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;
- the ability to take steps, such as disengagement or deactivation of systems, when such systems demonstrate unintended behaviour.

Metrics: AI Governability (Data Governance provides a foundation for AI Governance) depends on multiple other properties of AI, so different metrics should be applied in order to measure and guarantee the Governability. An example is the following:

1. Data lineage: Tracking compliance with data origin, flow and processing rules.
2. Data quality: Measuring the accuracy, relevance and completeness of data.
3. Compliance with AI ethics guidelines: Monitoring the percentage of projects adhering to established ethical guidelines.
4. AI system downtime and reliability: Tracking system uptime, response times and failure rates.
5. Security incidents: Monitoring the number of breach attempts or data exposure incidents.
6. Incident response time: Understanding how long it takes to identify, respond and mitigate AI-related incidents.
7. Stakeholder satisfaction and feedback: Using surveys to assess transparency and accountability of AI systems.

Interpretability [TAID-18]

References: Assessment List for Trustworthy Artificial Intelligence.

Definition: Interpretability refers to the concept of comprehensibility, explainability, or understandability. When an element of an AI system is interpretable, this means that it is possible, at least for an external observer, to understand it and find its meaning.

METRICS: See LIME and COVAR methods for AI Interpretability.

Recoverability [TAID-43]

REFERENCES: ISO/IEC 25010:2023

DEFINITION: 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.

METRICS: DORA metrics could be used or some of them taken as an example of meaningful metrics for Recoverability.

Responsibility [TAID-26]

REFERENCES: ISO/IEC 38500:2015

DEFINITION: Obligation to act and take decisions to achieve required outcomes.

METRICS: Refer to Transparency property to evaluate the decision-making process of an AI system.

Traceability [TAID-35]

REFERENCES: Adapted by EICACS from Ethics Guidelines for Trustworthy AI.

DEFINITION: The capability to track system data and events during the development, deployment, operation processes, and decommission.

METRICS: Traceability can be measured in different ways along the lifecycle of the AI system, for example:

- Perform risk assessment throughout the whole lifecycle.
- Provide documentation for each phase.
- Implement continuous quality control.
- Implement periodic auditing on the system to identify concept drifts or changes.
- Implement AI logging.
- Implement human oversight.[10]

[10] Traceability (future-ai.eu)

2. Security – CIA properties

Confidentiality [TAID-09]

REFERENCES: ISO/IEC 27000:2018

DEFINITION: Information is not made available or disclosed to unauthorized individuals, entities, or processes.

METRICS: Measuring the confidentiality of an AI system involves several aspects, including data security, privacy, ethics, and regulatory compliance.[11]

Data Integrity [TAID-17]

REFERENCES: ISO/IEC 27000:2018 and EASA Concept Paper: first usable guidance for Level 1 & 2 machine learning.

DEFINITION: It refers to the assurance that data and its values remain unaltered and uncorrupted throughout the processes of collection, storage, and processing.

METRICS: More than one technique:

- Checksums and Hashing: Like in traditional software, AI models can have a checksum or hash value calculated post-training. Before each execution, the current model's hash can be recalculated and compared to the original.
- Watermarking: Implanting unique signatures or watermarks into models. These watermarks can then be checked to validate the model's authenticity.
- Runtime Behavior Analysis: By monitoring the runtime behavior of models, any anomalies or deviations can signal potential integrity breaches.
- Provenance Tracking: Maintain a detailed log of all the model's interactions, updates, and changes. This not only helps in verification but also in tracing back any possible compromises.[12]

Availability [TAID-07]

REFERENCES: ISO/IEC 27000:2018

DEFINITION: Being accessible and usable on demand by an authorized entity.

[11] https://doi.org/10.3390/e25101429

[12] Model Integrity Verification: The Essential Guide | Nightfall AI Security 101

METRICS: Availability can be measured with the formula.

Availability = MTTF / (MTTF + MTTR), where:

MTTF = Mean time to failure

MTTR = Mean time to recovery

3. Safety Properties

AI self-protection [TAID-04]

REFERENCE: None.

DEFINITION: Integrated features and increased capacity of the AI to prevent non-intended third-party interactions like disclosure, reverse engineering, and miss-usage.

METRICS: None.

Autonomy [TAID-06]

REFERENCES: None.

DEFINITION: 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.

METRICS: None.

Autonomy (level of) for weapon systems

REFERENCES: Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues Daniele Amoroso & Guglielmo Tamburrini.

DEFINITION: Level of autonomy could be defined in the following way as 5 different levels.

- L1. A human engages with and selects targets and initiates any attack.
- L2. A program suggests alternative targets and a human chooses which to attack.
- L3. A program selects targets, and a human must approve before the attack.
- L4. A program selects and engages targets but is supervised by a human who retains the power to override its choices and abort the attack.
- L5: A program selects targets and initiates an attack on the basis of the mission goals as defined at the planning/activation stage, without further human involvement.

METRICS: None.

Homologation/Certification [TAID-16]

REFERENCE: None.

DEFINITION: The processes followed to homologate or certify a system as preconditions to release it for operation.

METRICS: None.

Sovereignty [TAID-29]

REFERENCE: None.

DEFINITION: The deployment of technology encourages or ensures proper sovereignty for state members, and/or EU.

METRICS: None.

4. AI Development Properties

Data Management Level (Datasets properties)

Data Completeness [TAID-08]

REFERENCES: EASA Concept Paper: first usable guidance for Level 1 & 2 machine learning applications

DEFINITION: Degree to which a data set sufficiently (according to specified criteria) covers the operational design domain for the intended application.

METRICS: Various techniques:

- Null check: find and fill empty or null data points in the dataset.
- Coverage check: make sure your data covers all necessary dimensions of the entity it represents.
- Missing value analysis: identify patterns in missing data to find systematic data collection issues.
- Data imputation: fill in missing data based on various strategies like mean, median, mode, or predictive modeling.
- Cross-reference check: compare your data with a trusted source to identify any missing elements.
- Cardinality check: assess if the number of unique values in a field matches expectations.
- Data sufficiency verification: ensure you have enough data to support your analysis and conclusions.
- Business rule confirmation: verify that all business rules or conditions are met in the data collection process.

Consistency [TAID-10]

REFERENCES: ISO/IEC 25012:2008

DEFINITION: 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.

METRICS: Various techniques:

- Cross-system check: compare data across different systems. They should match.
- Standardization: maintain uniform data formats. For instance, date fields should follow one format throughout.
- Data deduplication: remove duplicate data entries to avoid confusion and inconsistency.

- Business rule check: ensure data complies with the rules or constraints defined by your business requirements.
- Harmonization: align disparate data representations to achieve uniformity.
- Entity resolution: identify and link different representations of the same entity within or across datasets.
- Temporal consistency check: check if data maintains logical order and sequencing over time.[13]

Function gain, extension [TAID-13]

REFERENCES: None.

DEFINITION: The usage of the AI component/technology produces a function gain, or an extension of existing function(s) originally developed without AI.

METRICS: None.

Representability [TAID-24]

REFERENCES: EASA Concept Paper - first usable guidance for Level 1 & 2 machine learning applications.

DEFINITION: A data set is representative when the distribution of its key characteristics is similar to the actual input state space for the intended application.

Training outside the boundaries must be considered.

METRICS: A generic representativeness verification method is viewed as function (D) taking as input data sets and returning a probability of them being in-distribution.

Two opposite requirements must then hold:

(1) The probability of D evaluated on in-distribution data sets is high.

(2) The probability of D evaluated on out-of-distribution data sets is low.

The exact verification setting is to be determined depending on the required statistical significance and use case, but the framework remains method- and data-agnostic. Moreover, it is meant to allow easy verification as only in- or out-of-distribution (unannotated) data is required.

[13] Data Quality Metrics for Integrity, Consistency, and Compliance (atlan.com)

Observability [TAID-38]

REFERENCES: None.

DEFINITION: Observability is a measure of how well internal states of a system can be inferred from knowledge of its external outputs.

METRICS: None.

Bias [TAID-44]

REFERENCES: ISO/IEC TR 24027:2021

DEFINITION: Systematic difference in treatment of certain objects, people or groups in comparison to others.

METRICS: Use a statistical approach.[14]

Data Balance [TAID-45]

REFERENCES: H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

DEFINITION: In a balanced dataset, each class contributes equally to the overall composition. On the other hand, unbalanced datasets present underrepresented data, which may introduce bias towards the overrepresented classes.

METRICS: The accuracy of a classifier is the total number of correct predictions by the classifier divided by the total number of predictions. This may be good enough for a well-balanced class but not ideal for the imbalanced class problem. The other metrics such as precision is the measure of how accurate the classifier's prediction of a specific class and recall is the measure of the classifier's ability to identify a class.

For an imbalanced class dataset F1 score is a more appropriate metric.[15]

Data accuracy

Data Currentness

[14] https://arxiv.org/abs/2304.13680v2

[15] What is Imbalanced Data | Techniques to Handle Imbalanced Data (analyticsvidhya.com)

REFERENCES: A. Steimers and M. Schneider. 2022. Sources of Risk of AI Systems. International Journal of Environmental Research and Public Health 19, no. 6: 3641. https://doi.org/10.3390/ijerph19063641

DEFINITION: Currentness is the extent to which data has attributes that are the correct age in a particular context of use. Data that is current provides more accurate insights.

METRICS: Provide timestamp to identify when the data was collected, modified and if it has a baseline period.

Data Timeliness [TAID-34]

REFERENCES: Steimers A, Schneider M. Sources of Risk of AI Systems. International Journal of Environmental Research and Public Health. 2022; 19(6):3641. https://doi.org/10.3390/ijerph19063641

DEFINITION: Timeliness indicates the extent to which data from a source arrives quickly enough to be relevant. Timeliness refers to the latency between the time that a phenomenon occurs and the time the data recorded for that phenomenon are available for use; this dimension of data quality is particularly important when the dataset is a continuous stream of data.

METRICS: Timeliness is a measure that tells us about the lag between something that happened and when it was recorded. When we apply timeliness to data, we come up with the concept of data timeliness. Data timeliness uses the most recent timestamps in the dataset to calculate the time lag.

The formula that calculates data timeliness compares two timestamps and measures the time difference between them: the data lag.

The first value is the most recent timestamp in the dataset. It is the point in time when SOMETHING happened. It can be a business action, such as the timestamp of the most recent transaction in an eCommerce platform, the most recent impression of an advertisement, the timestamp of a log entry, or the timestamp of the last shipment.

The second value is the system's current time. It is the timestamp when we recorded the state of the data in the dataset.

The only challenge here is to identify the right timestamp column in the dataset that can tell us about the currency of the data.[16]

[16] How to measure data timeliness, freshness and staleness metrics (dqops.com)

5. Model Engineering and Development Properties

AI Performance metrics

Accountability [TAID-01]

REFERENCES: ISO/IEC 22989:2022

DEFINITION: State of being answerable for actions, decisions and performance.

METRICS: See related research.[17]

Accuracy [TAID-02]

REFERENCES: Steimers A, Schneider M. Sources of Risk of AI Systems. *International Journal of Environmental Research and Public Health*. 2022; 19(6):3641. https://doi.org/10.3390/ijerph19063641

DEFINITION: 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.

METRICS: use accuracy formula and confusion matrix.[18]

Generalisation [TAID-14]

REFERENCES: EASA Concept Paper: first usable guidance for Level 1 & 2 machine learning applications

DEFINITION: Generalization is the ability of ML models to provide accurate outputs when fed with inputs not seen during the training phase, which means that the "in-sample errors" should be a good approximation of the "out-of-sample" errors.

METRICS: There are various metrics for example:

- Measure performance on different data sets.[19]

---

[17] [Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, and Zhenchang Xing. 2024. Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability. In Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI (CAIN '24). Association for Computing Machinery, New York, NY, USA, 100–111. https://doi.org/10.1145/3644815.3644959].

[18] Accuracy score | CloudFactory Computer Vision Wiki

[19] Grosse, Roger. "Lecture 9: Generalization." (2018).

- Use the Inductive Bias Complexity Measure.[20]
- Evaluate system output after input augmentation.[21]

Predictability [TAID-20]

REFERENCES: None**.**

DEFINITION: Property of an AI system that enables reliable assumptions by stakeholders about the output.

METRICS: None**.**

Recognition [TAID-21]

REFERENCES: Pattern_Recognition_and_Machine_Learning**.**

DEFINITION: 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.

METRICS: None**.**

Reusability [TAID-27]

REFERENCES: None**.**

DEFINITION: Increased possibilities for reusage of the AI technology (or the system that integrates it) under larger/new operational conditions.

METRICS: None**.**

Specifiability [TAID-30]

REFERENCES: White Paper Machine Learning in Certified Systems.

DEFINITION: Extent to which the AI constituent can be correctly and completely described through a list of requirements.

---

[20] "Model-agnostic Measure of Generalization Difficulty." undefined (2023). doi: 10.48550/arxiv.2305.01034

[21] Sumukh, Aithal, K., Dhruva, Kashyap., Natarajan, Subramanyam. "Robustness to Augmentations as a Generalization metric.." arXiv: Learning, undefined (2021).

METRICS: Can be evaluated by the following metrics:

- Representability: degree of requested functions representation.
- Correctness: degree of correctness of described specifications and their formats.
- Non-restrictiveness: degree of descriptive format non-restriction.
- Ease of Description: degree of description ease.
- Ease of Modification: degree of modification or appendment ease.
- Described steps: actually described steps.[22]

Stability [TAID-31]

REFERENCES: EASA Concept Paper: first usable guidance for Level 1 & 2 machine learning applications

DEFINITION: Stability of the learning algorithm refers to ensuring that the produced model does not change a lot under perturbations of the training data set.

Stability of the model refers to keeping input-output relations of the model under small perturbations.

METRICS: Use PSI (Population Stability Index).

Usability [TAID-37]

REFERENCES: None**.**

DEFINITION: Increased possibilities for the usage of the AI technology (or the system that integrates it) under predefined operational conditions.

METRICS: None**.**

Quality [TAID-39]

REFERENCES: OECD HANDBOOK FOR INTERNATIONALLY COMPARATIVE EDUCATION STATISTICS.

[22] T. Miyoshi, Y. Togashi and M. Azuma, "Evaluating software development environment quality," [1989] Proceedings of the Thirteenth Annual International Computer Software & Applications Conference, Orlando, FL, USA, 1989, pp. 501-508, doi: 10.1109/CMPSAC.1989.65134. keywords: {Programming;Software quality;Software tools;Software prototyping;Software engineering;Standards development;ISO standards;IEC standards;Software measurement;Software design}

DEFINITION: "Fitness for use" for users' needs. 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.

METRICS: None**.**

Causality [TAID-40]

REFERENCES: None**.**

DEFINITION: Ability to establish causal relationship between events to ensure the fair behaviour of systems.

METRICS: None**.**